(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 October 2002 (03.10.2002)

PCT

(10) International Publication Number
WO 02/077234 A2

(51) International Patent Classification[7]: C12N 15/12, C07K 14/47, C12Q 1/68, C07K 16/18, G01N 33/50, A61K 38/17

(21) International Application Number: PCT/US01/48414

(22) International Filing Date: 31 October 2001 (31.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/244,758 31 October 2000 (31.10.2000) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/244,758 (CIP)
Filed on 31 October 2000 (31.10.2000)

(71) Applicant *(for all designated States except US)*: **DIADEXUS, INC.** [US/US]; 343 Oyster Point Boulevard, South San Francisco, CA 94080 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **SUN, Yongming** [CN/US]; 551 Shoal Drive, Redwood City, CA 94065 (US). **RECIPON, Herve** [FR/US]; 85 Fortuna Avenue, San Francisco, CA 94115 (US). **GHOSH, Malavika, G.** [IN/US]; 1733 Via Lugano, San Jose, CA 95120 (US). **LIU, Chenghua** [CN/US]; 1125 Ranchero Way #14, San Jose, CA 95117 (US).

(74) Agents: **LICATA, Jane, Massey** et al.; Licata & Tyrrel P.C., 66 E. Main Street, Marlton, NJ 08053 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/077234 A2

(54) Title: COMPOSITIONS AND METHODS RELATING TO COLON SPECIFIC GENES AND PROTEINS

(57) Abstract: The present invention relates to newly identified nucleic acids and polypeptides present in normal and neoplastic colon cells, including fragments, variants and derivatives of the nucleic acids and polypeptides. The present invention also relates to antibodies to the polypeptides of the invention, as well as agonists and antagonists of the polypeptides of the invention. The invention also relates to compositions comprising the nucleic acids, polypeptides, antibodies, variants, derivatives, agonists and antagonists of the invention and methods for the use of these compositions. These uses include identifying, diagnosing, monitoring, staging, imaging and treating colon cancer and non cancerous disease states in colon tissue, identifying colon tissue, monitoring and identifying and/or designing agonists and antagonists of polypeptides of the invention. The uses also include gene therapy, production of transgenic animals and cells, and production of engineered colon tissue for treatment and research.

COMPOSITIONS AND METHODS RELATING TO COLON SPECIFIC GENES AND PROTEINS

This application claims the benefit of priority from U.S. Provisional Application Serial No. 60/244,758 filed October 31, 2000, which is herein incorporated by reference in its entirety.

FIELD OF THE INVENTION

The present invention relates to newly identified nucleic acid molecules and polypeptides present in normal and neoplastic colon cells, including fragments, variants and derivatives of the nucleic acids and polypeptides. The present invention also relates to antibodies to the polypeptides of the invention, as well as agonists and antagonists of the polypeptides of the invention. The invention also relates to compositions comprising the nucleic acids, polypeptides, antibodies, variants, derivatives, agonists and antagonists of the invention and methods for the use of these compositions. These uses include identifying, diagnosing, monitoring, staging, imaging and treating colon cancer and non-cancerous disease states in colon tissue, identifying colon tissue and monitoring and identifying and/or designing agonists and antagonists of polypeptides of the invention. The uses also include gene therapy, production of transgenic animals and cells, and production of engineered colon tissue for treatment and research.

BACKGROUND OF THE INVENTION

Colorectal cancer is the second most common cause of cancer death in the United States and the third most prevalent cancer in both men and women. M. L. Davila & A. D. Davila, *Screening for Colon and Rectal Cancer, in* Colon and Rectal Cancer 47 (Peter S. Edelstein ed., 2000). Approximately 100,000 patients every year suffer from colon cancer and approximately half that number die of the disease. Hannah-Ngoc Ha & Bard C. Cosman, *Treatment of Colon Cancer, in* Colon and Rectal Cancer 157 (Peter S. Edelstein ed., 2000). Nearly all cases of colorectal cancer arise from adenomatous polyps, some of which mature into large polyps, undergo abnormal growth and development, and ultimately progress into cancer. Davila & Davila, *supra* at 55-56. This progression would appear to take at least 10 years in most patients, rendering it a

readily treatable form of cancer if diagnosed early, when the cancer is localized. *Id.* at 56; Walter J. Burdette, Cancer: Etiology, Diagnosis, and Treatment 125 (1998).

Although our understanding of the etiology of colon cancer is undergoing continual refinement, extensive research in this area points to a combination of factors, including age, hereditary and nonheriditary conditions, and environmental/dietary factors. Age is a key risk factor in the development of colorectal cancer, Davila & Davila, *supra* at 48, with men and women over 40 years of age become increasingly susceptible to that cancer, Burdette, *supra* at 126. Incidence rates increase considerably in each subsequent decade of life. Davila et al., *supra* at 48. A number of hereditary and nonhereditary conditions have also been linked to a heightened risk of developing colorectal cancer, including familial adenomatous polyposis (FAP), hereditary nonpolyposis colorectal cancer (Lynch syndrome or HNPCC), a personal and/or family history of colorectal cancer or adenomatous polyps, inflammatory bowel disease, diabetes mellitus, and obesity. *Id.* at 47; Henry T. Lynch & Jane F. Lynch, *Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndromes), in* Colon and Rectal Cancer 67-68 (Peter S. Edelstein ed., 2000).

In the case of FAP, the tumor suppressor gene APC (adenomatous polyposis coli), located at 5q21, has been either mutationally inactivated or deleted. Alberts et al., Molecular Biology of the Cell 1288 (3d ed. 1994). The APC protein plays a role in a number of functions, including cell adhesion, apoptosis, and repression of the *c-myc* oncogene. N. R. Hall & R. D. Madoff, *Genetics and the Polyp-Cancer Sequence*, Colon and Rectal Cancer 8 (Peter S. Edelstein, ed., 2000). Of those patients with colorectal cancer who have normal APC genes, over 65% have such mutations in the cancer cells but not in other tissues. Alberts et al., *supra* at 1288. In the case of HPNCC, patients manifest abnormalities in the tumor suppressor gene HNPCC, but only about 15% of tumors contain the mutated gene. *Id.* A host of other genes have also been implicated in colorectal cancer, including the K-*ras,* N-*ras,* H-*ras* and c-*myc* oncogenes, and the tumor suppressor genes *DCC* (deleted in colon carcinoma) and *p53*. Hall & Madoff, *supra* at 8-9; Alberts et al., *supra* at 1288.

Environmental/dietary factors associated with an increased risk of colorectal cancer include a high fat diet, intake of high dietary red meat, and sedentary lifestyle. Davila & Davila, *supra* at 47; Reddy, B. S., *Prev. Med.* 16(4): 460-7 (1987).

Conversely, environmental/dietary factors associated with a reduced risk of colorectal cancer include a diet high in fiber, folic acid, calcium, and hormone-replacement therapy in post-menopausal women. Davila & Davila, *supra* at 50-55. The effect of antioxidants in reducing the risk of colon cancer is unclear. *Id.* at 53.

Because colon cancer is highly treatable when detected at an early, localized stage, screening should be a part of routine care for all adults starting at age 50, especially those with first-degree relatives with colorectal cancer. One major advantage of colorectal cancer screening over its counterparts in other types of cancer is its ability to not only detect precancerous lesions, but to remove them as well. Davila & Davila, *supra* at 56. The key colorectal cancer screening tests in use today are fecal occult blood test, sigmoidoscopy, colonoscopy, double-contrast barium enema, and the carcinoembryonic antigen (CEA) test. *Id;* Burdette, *supra* at 125.

The fecal occult blood test (FOBT) screens for colorectal cancer by detecting the amount of blood in the stool, the premise being that neoplastic tissue, particularly malignant tissue, bleeds more than typical mucosa, with the amount of bleeding increasing with polyp size and cancer stage. Davila & Davila, *supra* at 56-57. While effective at detecting early stage tumors, FOBT is unable to detect adenomatous polyps (premalignant lesions), and, depending on the contents of the fecal sample, is subject to rendering false positives. *Id.* at 56-59. Sigmoidoscopy and colonoscopy, by contrast, allow direct visualization of the bowel, and enable one to detect, biopsy, and remove adenomatous polyps. *Id.* at 59-60, 61. Despite the advantages of these procedures, there are accompanying downsides: sigmoidoscopy, by definition, is limited to the sigmoid colon and below, colonoscopy is a relatively expensive procedure, and both share the risk of possible bowel perforation and hemorrhaging. *Id.* at 59-60. Double-contrast barium enema (DCBE) enables detection of lesions better than FOBT, and almost as well a colonoscopy, but it may be limited in evaluating the winding rectosigmoid region. *Id.* at 60. The CEA blood test, which involves screening the blood for carcinoembryonic antigen, shares the downside of FOBT, in that it is of limited utility in detecting colorectal cancer at an early stage. Burdette, *supra* at 125.

Once colon cancer has been diagnosed, treatment decisions are typically made in reference to the stage of cancer progression. A number of techniques are employed to stage the cancer (some of which are also used to screen for colon cancer), including

pathologic examination of resected colon, sigmoidoscopy, colonoscopy, and various imaging techniques. AJCC Cancer Staging Handbook 84 (Irvin D. Fleming et al. eds., 5th ed. 1998); Montgomery, R. C. and Ridge, J.A., *Semin. Surg. Oncol.* 15(3): 143-150 (1998). Moreover, chest films, liver functionality tests, and liver scans are employed to determine the extent of metastasis. Fleming et al. eds., *supra* at 84. While computerized tomography and magnetic resonance imaging are useful in staging colorectal cancer in its later stages, both have unacceptably low staging accuracy for identifying early stages of the disease, due to the difficulty that both methods have in (1) revealing the depth of bowel wall tumor infiltration and (2) diagnosing malignant adenopathy. Thoeni, R. F., *Radiol. Clin. N. Am.* 35(2): 457-85 (1997). Rather, techniques such as transrectal ultrasound (TRUS) are preferred in this context, although this technique is inaccurate with respect to detecting small lymph nodes that may contain metastases. David Blumberg & Frank G. Opelka, *Neoadjuvant and Adjuvant Therapy for Adenocarcinoma of the Rectum, in* Colon and Rectal Cancer 316 (Peter S. Edelstein ed., 2000).

Several classification systems have been devised to stage the extent of colorectal cancer, including the Dukes' system and the more detailed International Union against Cancer-American Joint Committee on Cancer TNM staging system, which is considered by many in the field to be a more useful staging system. Burdette, *supra* at 126-27. The TNM system, which is used for either clinical or pathological staging, is divided into four stages, each of which evaluates the extent of cancer growth with respect to primary tumor (T), regional lymph nodes (N), and distant metastasis (M). Fleming et al. eds., *supra* at 84-85. The system focuses on the extent of tumor invasion into the intestinal wall, invasion of adjacent structures, the number of regional lymph nodes that have been affected, and whether distant metastasis has occurred. *Id.* at 81.

Stage 0 is characterized by *in situ* carcinoma (Tis), in which the cancer cells are located inside the glandular basement membrane (intraepithelial) or lamina propria (intramucosal). *Id.* at 84-85; Burdette, *supra* at 127. In this stage, the cancer has not spread to the regional lymph nodes (N0), and there is no distant metastasis (M0). Fleming et al. eds., *supra* at 85; Burdette, *supra* at 127. In stage I, there is still no spread of the cancer to the regional lymph nodes and no distant metastasis, but the tumor has invaded the submucosa (T1) or has progressed further to invade the muscularis propria (T2). Fleming et al. eds., *supra* at 84-85; Burdette, *supra* at 127. Stage II also involves

no spread of the cancer to the regional lymph nodes and no distant metastasis, but the tumor has invaded the subserosa, or the nonperitonealized pericolic or perirectal tissues (T3), or has progressed to invade other organs or structures, and/or has perforated the visceral peritoneum (T4). *Id.* Stage 3 is characterized by any of the T substages, no distant metastasis, and either metastasis in 1 to 3 regional lymph nodes (N1) or metastasis in four or more regional lymph nodes (N2). Fleming et al. eds., *supra* at 85; Burdette, *supra* at 127. Lastly, stage 4 involves any of the T or N substages, as well as distant metastasis. *Id.*

Currently, pathological staging of colon cancer is preferable over clinical staging as pathological staging provides a more accurate prognosis. Pathological staging typically involves examination of the resected colon section, along with surgical examination of the abdominal cavity. Fleming et al. eds., *supra* at 84. Clinical staging would be a preferred method of staging were it at least as accurate as pathological staging, as it does not depend on the invasive procedures of its counterpart.

Turning to the treatment of colorectal cancer, surgical resection results in a cure for roughly 50% of patients. Burdette, *supra* at125. Irradiation is used both preoperatively and postoperatively in treating colorectal cancer. *Id.* at 125, 132-33. Chemotherapeutic agents, particularly 5-fluorouracil, are also powerful weapons in treating colorectal cancer. *Id.* at 125, 133. Other agents include irinotecan and floxuridine, cisplatin, levamisole, methotrexate, interferon-alpha, and leucovorin. *Id.* at 133. Nonetheless, thirty to forty percent of patients will develop a recurrence of colon cancer following surgical resection. Wayne De Vos, *Follow-up After Treatment of Colon Cancer*, Colon and Rectal Cancer 225 (Peter S. Edelstein ed., 2000), which in many patients is the ultimate cause of death. Accordingly, colon cancer patients must be closely monitored to determine response to therapy and to detect persistent or recurrent disease and metastasis.

From the foregoing, it is clear that procedures used for detecting, diagnosing, monitoring, staging, prognosticating, and preventing the recurrence of colorectal cancer are of critical importance to the outcome of the patient. Moreover, current procedures, while helpful in each of these analyses, are limited by their specificity, sensitivity, invasiveness, and/or their cost. As such, highly specific and sensitive procedures that

would operate by way of detecting novel markers in cells, tissues, or bodily fluids, with minimal invasiveness and at a reasonable cost, would be highly desirable.

Accordingly, there is a great need for more sensitive and accurate methods for predicting whether a person is likely to develop colorectal cancer, for diagnosing colorectal cancer, for monitoring the progression of the disease, for staging the colorectal cancer, for determining whether the colorectal cancer has metastasized, and for imaging the colorectal cancer. There is also a need for better treatment of colorectal cancer.

SUMMARY OF THE INVENTION

The present invention solves these and other needs in the art by providing nucleic acid molecules and polypeptides as well as antibodies, agonists and antagonists, thereto that may be used to identify, diagnose, monitor, stage, image and treat colon cancer and non-cancerous disease states in colon; identify and monitor colon tissue; and identify and design agonists and antagonists of polypeptides of the invention. The invention also provides gene therapy, methods for producing transgenic animals and cells, and methods for producing engineered colon tissue for treatment and research.

Accordingly, one object of the invention is to provide nucleic acid molecules that are specific to colon cells and/or colon tissue. These colon specific nucleic acids (CSNAs) may be a naturally-occurring cDNA, genomic DNA, RNA, or a fragment of one of these nucleic acids, or may be a non-naturally-occurring nucleic acid molecule. If the CSNA is genomic DNA, then the CSNA is a colon specific gene (CSG). In a preferred embodiment, the nucleic acid molecule encodes a polypeptide that is specific to colon. In a more preferred embodiment, the nucleic acid molecule encodes a polypeptide that comprises an amino acid sequence of SEQ ID NO: 101 through 176. In another highly preferred embodiment, the nucleic acid molecule comprises a nucleic acid sequence of SEQ ID NO: 1 through 100. By nucleic acid molecule, it is also meant to be inclusive of sequences that selectively hybridize or exhibit substantial sequence similarity to a nucleic acid molecule encoding a CSP, or that selectively hybridize or exhibit substantial sequence similarity to a CSNA, as well as allelic variants of a nucleic acid molecule encoding a CSP, and allelic variants of a CSNA. Nucleic acid molecules comprising a part of a nucleic acid sequence that encodes a CSP or that comprises a part of a nucleic acid sequence of a CSNA are also provided.

A related object of the present invention is to provide a nucleic acid molecule comprising one or more expression control sequences controlling the transcription and/or translation of all or a part of a CSNA. In a preferred embodiment, the nucleic acid molecule comprises one or more expression control sequences controlling the transcription and/or translation of a nucleic acid molecule that encodes all or a fragment of a CSP.

Another object of the invention is to provide vectors and/or host cells comprising a nucleic acid molecule of the instant invention. In a preferred embodiment, the nucleic acid molecule encodes all or a fragment of a CSP. In another preferred embodiment, the nucleic acid molecule comprises all or a part of a CSNA.

Another object of the invention is to provided methods for using the vectors and host cells comprising a nucleic acid molecule of the instant invention to recombinantly produce polypeptides of the invention.

Another object of the invention is to provide a polypeptide encoded by a nucleic acid molecule of the invention. In a preferred embodiment, the polypeptide is a CSP. The polypeptide may comprise either a fragment or a full-length protein as well as a mutant protein (mutein), fusion protein, homologous protein or a polypeptide encoded by an allelic variant of a CSP.

Another object of the invention is to provide an antibody that specifically binds to a polypeptide of the instant invention..

Another object of the invention is to provide agonists and antagonists of the nucleic acid molecules and polypeptides of the instant invention.

Another object of the invention is to provide methods for using the nucleic acid molecules to detect or amplify nucleic acid molecules that have similar or identical nucleic acid sequences compared to the nucleic acid molecules described herein. In a preferred embodiment, the invention provides methods of using the nucleic acid molecules of the invention for identifying, diagnosing, monitoring, staging, imaging and treating colon cancer and non-cancerous disease states in colon. In another preferred embodiment, the invention provides methods of using the nucleic acid molecules of the invention for identifying and/or monitoring colon tissue. The nucleic acid molecules of the instant invention may also be used in gene therapy, for producing transgenic animals and cells, and for producing engineered colon tissue for treatment and research.

The polypeptides and/or antibodies of the instant invention may also be used to identify, diagnose, monitor, stage, image and treat colon cancer and non-cancerous disease states in colon. The invention provides methods of using the polypeptides of the invention to identify and/or monitor colon tissue, and to produce engineered colon tissue.

The agonists and antagonists of the instant invention may be used to treat colon cancer and non-cancerous disease states in colon and to produce engineered colon tissue.

Yet another object of the invention is to provide a computer readable means of storing the nucleic acid and amino acid sequences of the invention. The records of the computer readable means can be accessed for reading and displaying of sequences for comparison, alignment and ordering of the sequences of the invention to other sequences.

DETAILED DESCRIPTION OF THE INVENTION

Definitions and General Techniques

Unless otherwise defined herein, scientific and technical terms used in connection with the present invention shall have the meanings that are commonly understood by those of ordinary skill in the art. Further, unless otherwise required by context, singular terms shall include pluralities and plural terms shall include the singular. Generally, nomenclatures used in connection with, and techniques of, cell and tissue culture, molecular biology, immunology, microbiology, genetics and protein and nucleic acid chemistry and hybridization described herein are those well-known and commonly used in the art. The methods and techniques of the present invention are generally performed according to conventional methods well-known in the art and as described in various general and more specific references that are cited and discussed throughout the present specification unless otherwise indicated. *See, e.g.*, Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, 2d ed., Cold Spring Harbor Laboratory Press (1989) and Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, 3d ed., Cold Spring Harbor Press (2001); Ausubel *et al.*, Current Protocols in Molecular Biology, Greene Publishing Associates (1992, and Supplements to 2000); Ausubel *et al.*, Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology – 4th Ed., Wiley & Sons (1999); Harlow and Lane, Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press (1990); and Harlow and Lane, Using Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press (1999); each of which is incorporated herein by reference in its entirety.

Enzymatic reactions and purification techniques are performed according to manufacturer's specifications, as commonly accomplished in the art or as described herein. The nomenclatures used in connection with, and the laboratory procedures and techniques of, analytical chemistry, synthetic organic chemistry, and medicinal and pharmaceutical chemistry described herein are those well-known and commonly used in the art. Standard techniques are used for chemical syntheses, chemical analyses, pharmaceutical preparation, formulation, and delivery, and treatment of patients.

The following terms, unless otherwise indicated, shall be understood to have the following meanings:

A "nucleic acid molecule" of this invention refers to a polymeric form of nucleotides and includes both sense and antisense strands of RNA, cDNA, genomic DNA, and synthetic forms and mixed polymers of the above. A nucleotide refers to a ribonucleotide, deoxynucleotide or a modified form of either type of nucleotide. A "nucleic acid molecule" as used herein is synonymous with "nucleic acid" and "polynucleotide." The term "nucleic acid molecule" usually refers to a molecule of at least 10 bases in length, unless otherwise specified. The term includes single- and double-stranded forms of DNA. In addition, a polynucleotide may include either or both naturally-occurring and modified nucleotides linked together by naturally-occurring and/or non-naturally occurring nucleotide linkages.

The nucleic acid molecules may be modified chemically or biochemically or may contain non-natural or derivatized nucleotide bases, as will be readily appreciated by those of skill in the art. Such modifications include, for example, labels, methylation, substitution of one or more of the naturally occurring nucleotides with an analog, internucleotide modifications such as uncharged linkages (*e.g.*, methyl phosphonates, phosphotriesters, phosphoramidates, carbamates, etc.), charged linkages (*e.g.*, phosphorothioates, phosphorodithioates, etc.), pendent moieties (*e.g.*, polypeptides), intercalators (*e.g.*, acridine, psoralen, etc.), chelators, alkylators, and modified linkages (*e.g.*, alpha anomeric nucleic acids, etc.) The term "nucleic acid molecule" also includes any topological conformation, including single-stranded, double-stranded, partially duplexed, triplexed, hairpinned, circular and padlocked conformations. Also included are synthetic molecules that mimic polynucleotides in their ability to bind to a designated sequence via hydrogen bonding and other chemical interactions. Such molecules are

known in the art and include, for example, those in which peptide linkages substitute for phosphate linkages in the backbone of the molecule.

A "gene" is defined as a nucleic acid molecule that comprises a nucleic acid sequence that encodes a polypeptide and the expression control sequences that surround the nucleic acid sequence that encodes the polypeptide. For instance, a gene may comprise a promoter, one or more enhancers, a nucleic acid sequence that encodes a polypeptide, downstream regulatory sequences and, possibly, other nucleic acid sequences involved in regulation of the expression of an RNA. As is well-known in the art, eukaryotic genes usually contain both exons and introns. The term "exon" refers to a nucleic acid sequence found in genomic DNA that is bioinformatically predicted and/or experimentally confirmed to contribute a contiguous sequence to a mature mRNA transcript. The term "intron" refers to a nucleic acid sequence found in genomic DNA that is predicted and/or confirmed to not contribute to a mature mRNA transcript, but rather to be "spliced out" during processing of the transcript.

A nucleic acid molecule or polypeptide is "derived" from a particular species if the nucleic acid molecule or polypeptide has been isolated from the particular species, or if the nucleic acid molecule or polypeptide is homologous to a nucleic acid molecule or polypeptide isolated from a particular species.

An "isolated" or "substantially pure" nucleic acid or polynucleotide (*e.g.*, an RNA, DNA or a mixed polymer) is one which is substantially separated from other cellular components that naturally accompany the native polynucleotide in its natural host cell, *e.g.*, ribosomes, polymerases, or genomic sequences with which it is naturally associated. The term embraces a nucleic acid or polynucleotide that (1) has been removed from its naturally occurring environment, (2) is not associated with all or a portion of a polynucleotide in which the "isolated polynucleotide" is found in nature, (3) is operatively linked to a polynucleotide which it is not linked to in nature, (4) does not occur in nature as part of a larger sequence or (5) includes nucleotides or internucleoside bonds that are not found in nature. The term "isolated" or "substantially pure" also can be used in reference to recombinant or cloned DNA isolates, chemically synthesized polynucleotide analogs, or polynucleotide analogs that are biologically synthesized by heterologous systems. The term "isolated nucleic acid molecule" includes nucleic acid molecules that are integrated into a host cell chromosome at a heterologous site,

recombinant fusions of a native fragment to a heterologous sequence, recombinant vectors present as episomes or as integrated into a host cell chromosome.

A "part" of a nucleic acid molecule refers to a nucleic acid molecule that comprises a partial contiguous sequence of at least 10 bases of the reference nucleic acid molecule. Preferably, a part comprises at least 15 to 20 bases of a reference nucleic acid molecule. In theory, a nucleic acid sequence of 17 nucleotides is of sufficient length to occur at random less frequently than once in the three gigabase human genome, and thus to provide a nucleic acid probe that can uniquely identify the reference sequence in a nucleic acid mixture of genomic complexity. A preferred part is one that comprises a nucleic acid sequence that can encode at least 6 contiguous amino acid sequences (fragments of at least 18 nucleotides) because they are useful in directing the expression or synthesis of peptides that are useful in mapping the epitopes of the polypeptide encoded by the reference nucleic acid. *See, e.g.*, Geysen *et al.*, *Proc. Natl. Acad. Sci. USA* 81:3998-4002 (1984); and United States Patent Nos. 4,708,871 and 5,595,915, the disclosures of which are incorporated herein by reference in their entireties. A part may also comprise at least 25, 30, 35 or 40 nucleotides of a reference nucleic acid molecule, or at least 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400 or 500 nucleotides of a reference nucleic acid molecule. A part of a nucleic acid molecule may comprise no other nucleic acid sequences. Alternatively, a part of a nucleic acid may comprise other nucleic acid sequences from other nucleic acid molecules.

The term "oligonucleotide" refers to a nucleic acid molecule generally comprising a length of 200 bases or fewer. The term often refers to single-stranded deoxyribonucleotides, but it can refer as well to single- or double-stranded ribonucleotides, RNA:DNA hybrids and double-stranded DNAs, among others. Preferably, oligonucleotides are 10 to 60 bases in length and most preferably 12, 13, 14, 15, 16, 17, 18, 19 or 20 bases in length. Other preferred oligonucleotides are 25, 30, 35, 40, 45, 50, 55 or 60 bases in length. Oligonucleotides may be single-stranded, *e.g.* for use as probes or primers, or may be double-stranded, *e.g.* for use in the construction of a mutant gene. Oligonucleotides of the invention can be either sense or antisense oligonucleotides. An oligonucleotide can be derivatized or modified as discussed above for nucleic acid molecules.

Oligonucleotides, such as single-stranded DNA probe oligonucleotides, often are synthesized by chemical methods, such as those implemented on automated oligonucleotide synthesizers. However, oligonucleotides can be made by a variety of other methods, including *in vitro* recombinant DNA-mediated techniques and by expression of DNAs in cells and organisms. Initially, chemically synthesized DNAs typically are obtained without a 5' phosphate. The 5' ends of such oligonucleotides are not substrates for phosphodiester bond formation by ligation reactions that employ DNA ligases typically used to form recombinant DNA molecules. Where ligation of such oligonucleotides is desired, a phosphate can be added by standard techniques, such as those that employ a kinase and ATP. The 3' end of a chemically synthesized oligonucleotide generally has a free hydroxyl group and, in the presence of a ligase, such as T4 DNA ligase, readily will form a phosphodiester bond with a 5' phosphate of another polynucleotide, such as another oligonucleotide. As is well-known, this reaction can be prevented selectively, where desired, by removing the 5' phosphates of the other polynucleotide(s) prior to ligation.

The term "naturally-occurring nucleotide" referred to herein includes naturally-occurring deoxyribonucleotides and ribonucleotides. The term "modified nucleotides" referred to herein includes nucleotides with modified or substituted sugar groups and the like. The term "nucleotide linkages" referred to herein includes nucleotides linkages such as phosphorothioate, phosphorodithioate, phosphoroselenoate, phosphorodiselenoate, phosphoroanilothioate, phoshoraniladate, phosphoroamidate, and the like. *See e.g.*, LaPlanche *et al. Nucl. Acids Res.* 14:9081-9093 (1986); Stein *et al. Nucl. Acids Res.* 16:3209-3221 (1988); Zon *et al. Anti-Cancer Drug Design* 6:539-568 (1991); Zon *et al.*, in Eckstein (ed.) Oligonucleotides and Analogues: A Practical Approach, pp. 87-108, Oxford University Press (1991); United States Patent No. 5,151,510; Uhlmann and Peyman *Chemical Reviews* 90:543 (1990), the disclosures of which are hereby incorporated by reference.

Unless specified otherwise, the left hand end of a polynucleotide sequence in sense orientation is the 5' end and the right hand end of the sequence is the 3' end. In addition, the left hand direction of a polynucleotide sequence in sense orientation is referred to as the 5' direction, while the right hand direction of the polynucleotide sequence is referred to as the 3' direction. Further, unless otherwise indicated, each

nucleotide sequence is set forth herein as a sequence of deoxyribonucleotides. It is intended, however, that the given sequence be interpreted as would be appropriate to the polynucleotide composition: for example, if the isolated nucleic acid is composed of RNA, the given sequence intends ribonucleotides, with uridine substituted for thymidine.

The term "allelic variant" refers to one of two or more alternative naturally-occurring forms of a gene, wherein each gene possesses a unique nucleotide sequence. In a preferred embodiment, different alleles of a given gene have similar or identical biological properties.

The term "percent sequence identity" in the context of nucleic acid sequences refers to the residues in two sequences which are the same when aligned for maximum correspondence. The length of sequence identity comparison may be over a stretch of at least about nine nucleotides, usually at least about 20 nucleotides, more usually at least about 24 nucleotides, typically at least about 28 nucleotides, more typically at least about 32 nucleotides, and preferably at least about 36 or more nucleotides. There are a number of different algorithms known in the art which can be used to measure nucleotide sequence identity. For instance, polynucleotide sequences can be compared using FASTA, Gap or Bestfit, which are programs in Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, Wisconsin. FASTA, which includes, *e.g.,* the programs FASTA2 and FASTA3, provides alignments and percent sequence identity of the regions of the best overlap between the query and search sequences (Pearson, *Methods Enzymol.* 183: 63-98 (1990); Pearson, *Methods Mol. Biol.* 132: 185-219 (2000); Pearson, *Methods Enzymol.* 266: 227-258 (1996); Pearson, *J. Mol. Biol.* 276: 71-84 (1998); herein incorporated by reference). Unless otherwise specified, default parameters for a particular program or algorithm are used. For instance, percent sequence identity between nucleic acid sequences can be determined using FASTA with its default parameters (a word size of 6 and the NOPAM factor for the scoring matrix) or using Gap with its default parameters as provided in GCG Version 6.1, herein incorporated by reference.

A reference to a nucleic acid sequence encompasses its complement unless otherwise specified. Thus, a reference to a nucleic acid molecule having a particular sequence should be understood to encompass its complementary strand, with its

complementary sequence. The complementary strand is also useful, *e.g.*, for antisense therapy, hybridization probes and PCR primers.

In the molecular biology art, researchers use the terms "percent sequence identity", "percent sequence similarity" and "percent sequence homology" interchangeably. In this application, these terms shall have the same meaning with respect to nucleic acid sequences only.

The term "substantial similarity" or "substantial sequence similarity," when referring to a nucleic acid or fragment thereof, indicates that, when optimally aligned with appropriate nucleotide insertions or deletions with another nucleic acid (or its complementary strand), there is nucleotide sequence identity in at least about 50%, more preferably 60% of the nucleotide bases, usually at least about 70%, more usually at least about 80%, preferably at least about 90%, and more preferably at least about 95-98% of the nucleotide bases, as measured by any well-known algorithm of sequence identity, such as FASTA, BLAST or Gap, as discussed above.

Alternatively, substantial similarity exists when a nucleic acid or fragment thereof hybridizes to another nucleic acid, to a strand of another nucleic acid, or to the complementary strand thereof, under selective hybridization conditions. Typically, selective hybridization will occur when there is at least about 55% sequence identity, preferably at least about 65%, more preferably at least about 75%, and most preferably at least about 90% sequence identity, over a stretch of at least about 14 nucleotides, more preferably at least 17 nucleotides, even more preferably at least 20, 25, 30, 35, 40, 50, 60, 70, 80, 90 or 100 nucleotides.

Nucleic acid hybridization will be affected by such conditions as salt concentration, temperature, solvents, the base composition of the hybridizing species, length of the complementary regions, and the number of nucleotide base mismatches between the hybridizing nucleic acids, as will be readily appreciated by those skilled in the art. "Stringent hybridization conditions" and "stringent wash conditions" in the context of nucleic acid hybridization experiments depend upon a number of different physical parameters. The most important parameters include temperature of hybridization, base composition of the nucleic acids, salt concentration and length of the nucleic acid. One having ordinary skill in the art knows how to vary these parameters to achieve a particular stringency of hybridization. In general, "stringent hybridization" is

performed at about 25°C below the thermal melting point (T_m) for the specific DNA hybrid under a particular set of conditions. "Stringent washing" is performed at temperatures about 5°C lower than the T_m for the specific DNA hybrid under a particular set of conditions. The T_m is the temperature at which 50% of the target sequence hybridizes to a perfectly matched probe. *See* Sambrook (1989), *supra*, p. 9.51, hereby incorporated by reference.

The T_m for a particular DNA-DNA hybrid can be estimated by the formula:
T_m = 81.5°C + 16.6 ($\log_{10}[Na^+]$) + 0.41 (fraction G + C) – 0.63 (% formamide) - (600/l)
where l is the length of the hybrid in base pairs.

The T_m for a particular RNA-RNA hybrid can be estimated by the formula:
T_m = 79.8°C + 18.5 ($\log_{10}[Na^+]$) + 0.58 (fraction G + C) + 11.8 (fraction G + C)^2 - 0.35 (% formamide) - (820/l).

The T_m for a particular RNA-DNA hybrid can be estimated by the formula:
T_m = 79.8°C + 18.5($\log_{10}[Na^+]$) + 0.58 (fraction G + C) + 11.8 (fraction G + C)^2 - 0.50 (% formamide) - (820/l).

In general, the T_m decreases by 1-1.5°C for each 1% of mismatch between two nucleic acid sequences. Thus, one having ordinary skill in the art can alter hybridization and/or washing conditions to obtain sequences that have higher or lower degrees of sequence identity to the target nucleic acid. For instance, to obtain hybridizing nucleic acids that contain up to 10% mismatch from the target nucleic acid sequence, 10-15°C would be subtracted from the calculated T_m of a perfectly matched hybrid, and then the hybridization and washing temperatures adjusted accordingly. Probe sequences may also hybridize specifically to duplex DNA under certain conditions to form triplex or other higher order DNA complexes. The preparation of such probes and suitable hybridization conditions are well-known in the art.

An example of stringent hybridization conditions for hybridization of complementary nucleic acid sequences having more than 100 complementary residues on a filter in a Southern or Northern blot or for screening a library is 50% formamide/6X SSC at 42°C for at least ten hours and preferably overnight (approximately 16 hours). Another example of stringent hybridization conditions is 6X SSC at 68°C without formamide for at least ten hours and preferably overnight. An example of moderate stringency hybridization conditions is 6X SSC at 55°C without formamide for at least ten

hours and preferably overnight. An example of low stringency hybridization conditions for hybridization of complementary nucleic acid sequences having more than 100 complementary residues on a filter in a Southern or Northern blot or for screening a library is 6X SSC at 42°C for at least ten hours. Hybridization conditions to identify nucleic acid sequences that are similar but not identical can be identified by experimentally changing the hybridization temperature from 68°C to 42°C while keeping the salt concentration constant (6X SSC), or keeping the hybridization temperature and salt concentration constant (*e.g.* 42°C and 6X SSC) and varying the formamide concentration from 50% to 0%. Hybridization buffers may also include blocking agents to lower background. These agents are well-known in the art. *See* Sambrook *et al.* (1989), *supra*, pages 8.46 and 9.46-9.58, herein incorporated by reference. *See also* Ausubel (1992), *supra*, Ausubel (1999), *supra*, and Sambrook (2001), *supra*.

Wash conditions also can be altered to change stringency conditions. An example of stringent wash conditions is a 0.2x SSC wash at 65°C for 15 minutes (*see* Sambrook (1989), *supra*, for SSC buffer). Often the high stringency wash is preceded by a low stringency wash to remove excess probe. An exemplary medium stringency wash for duplex DNA of more than 100 base pairs is 1x SSC at 45°C for 15 minutes. An exemplary low stringency wash for such a duplex is 4x SSC at 40°C for 15 minutes. In general, signal-to-noise ratio of 2x or higher than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization.

As defined herein, nucleic acid molecules that do not hybridize to each other under stringent conditions are still substantially similar to one another if they encode polypeptides that are substantially identical to each other. This occurs, for example, when a nucleic acid molecule is created synthetically or recombinantly using high codon degeneracy as permitted by the redundancy of the genetic code.

Hybridization conditions for nucleic acid molecules that are shorter than 100 nucleotides in length (*e.g.*, for oligonucleotide probes) may be calculated by the formula:

$T_m = 81.5°C + 16.6(\log_{10}[Na^+]) + 0.41(\text{fraction G+C}) -(600/N)$,

wherein N is change length and the $[Na^+]$ is 1 M or less. *See* Sambrook (1989), *supra*, p. 11.46. For hybridization of probes shorter than 100 nucleotides, hybridization is usually performed under stringent conditions (5-10°C below the T_m) using high concentrations (0.1-1.0 pmol/ml) of probe. *Id.* at p. 11.45. Determination of hybridization using

mismatched probes, pools of degenerate probes or "guessmers," as well as hybridization solutions and methods for empirically determining hybridization conditions are well-known in the art. *See, e.g.*, Ausubel (1999), *supra*; Sambrook (1989), *supra*, pp. 11.45-11.57.

The term "digestion" or "digestion of DNA" refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes referred to herein are commercially available and their reaction conditions, cofactors and other requirements for use are known and routine to the skilled artisan. For analytical purposes, typically, 1 μg of plasmid or DNA fragment is digested with about 2 units of enzyme in about 20 μl of reaction buffer. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 μg of DNA are digested with 20 to 250 units of enzyme in proportionately larger volumes. Appropriate buffers and substrate amounts for particular restriction enzymes are described in standard laboratory manuals, such as those referenced below, and they are specified by commercial suppliers. Incubation times of about 1 hour at 37°C are ordinarily used, but conditions may vary in accordance with standard procedures, the supplier's instructions and the particulars of the reaction. After digestion, reactions may be analyzed, and fragments may be purified by electrophoresis through an agarose or polyacrylamide gel, using well-known methods that are routine for those skilled in the art.

The term "ligation" refers to the process of forming phosphodiester bonds between two or more polynucleotides, which most often are double-stranded DNAS. Techniques for ligation are well-known to the art and protocols for ligation are described in standard laboratory manuals and references, such as, *e.g.*, Sambrook (1989), *supra*.

Genome-derived "single exon probes," are probes that comprise at least part of an exon ("reference exon") and can hybridize detectably under high stringency conditions to transcript-derived nucleic acids that include the reference exon but do not hybridize detectably under high stringency conditions to nucleic acids that lack the reference exon. Single exon probes typically further comprise, contiguous to a first end of the exon portion, a first intronic and/or intergenic sequence that is identically contiguous to the exon in the genome, and may contain a second intronic and/or intergenic sequence that is identically contiguous to the exon in the genome. The minimum length of genome-

derived single exon probes is defined by the requirement that the exonic portion be of sufficient length to hybridize under high stringency conditions to transcript-derived nucleic acids, as discussed above. The maximum length of genome-derived single exon probes is defined by the requirement that the probes contain portions of no more than one exon. The single exon probes may contain priming sequences not found in contiguity with the rest of the probe sequence in the genome, which priming sequences are useful for PCR and other amplification-based technologies.

The term "microarray" or "nucleic acid microarray" refers to a substrate-bound collection of plural nucleic acids, hybridization to each of the plurality of bound nucleic acids being separately detectable. The substrate can be solid or porous, planar or non-planar, unitary or distributed. Microarrays or nucleic acid microarrays include all the devices so called in Schena (ed.), DNA Microarrays: A Practical Approach (Practical Approach Series), Oxford University Press (1999); *Nature Genet.* 21(1)(suppl.):1 - 60 (1999); Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000). These microarrays include substrate-bound collections of plural nucleic acids in which the plurality of nucleic acids are disposed on a plurality of beads, rather than on a unitary planar substrate, as is described, *inter alia*, in Brenner *et al.*, *Proc. Natl. Acad. Sci. USA* 97(4):1665-1670 (2000).

The term "mutated" when applied to nucleic acid molecules means that nucleotides in the nucleic acid sequence of the nucleic acid molecule may be inserted, deleted or changed compared to a reference nucleic acid sequence. A single alteration may be made at a locus (a point mutation) or multiple nucleotides may be inserted, deleted or changed at a single locus. In addition, one or more alterations may be made at any number of loci within a nucleic acid sequence. In a preferred embodiment, the nucleic acid molecule comprises the wild type nucleic acid sequence encoding a CSP or is a CSNA. The nucleic acid molecule may be mutated by any method known in the art including those mutagenesis techniques described *infra*.

The term "error-prone PCR" refers to a process for performing PCR under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. *See, e.g.*, Leung *et al.*, *Technique* 1: 11-15 (1989) and Caldwell *et al.*, *PCR Methods Applic.* 2: 28-33 (1992).

The term "oligonucleotide-directed mutagenesis" refers to a process which enables the generation of site-specific mutations in any cloned DNA segment of interest. *See, e.g.*, Reidhaar-Olson *et al.*, *Science* 241: 53-57 (1988).

The term "assembly PCR" refers to a process which involves the assembly of a PCR product from a mixture of small DNA fragments. A large number of different PCR reactions occur in parallel in the same vial, with the products of one reaction priming the products of another reaction.

The term "sexual PCR mutagenesis" or "DNA shuffling" refers to a method of error-prone PCR coupled with forced homologous recombination between DNA molecules of different but highly related DNA sequence *in vitro*, caused by random fragmentation of the DNA molecule based on sequence similarity, followed by fixation of the crossover by primer extension in an error-prone PCR reaction. *See, e.g.*, Stemmer, *Proc. Natl. Acad. Sci. U.S.A.* 91: 10747-10751 (1994). DNA shuffling can be carried out between several related genes ("Family shuffling").

The term "*in vivo* mutagenesis" refers to a process of generating random mutations in any cloned DNA of interest which involves the propagation of the DNA in a strain of bacteria such as *E. coli* that carries mutations in one or more of the DNA repair pathways. These "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA in a mutator strain will eventually generate random mutations within the DNA.

The term "cassette mutagenesis" refers to any process for replacing a small region of a double-stranded DNA molecule with a synthetic oligonucleotide "cassette" that differs from the native sequence. The oligonucleotide often contains completely and/or partially randomized native sequence.

The term "recursive ensemble mutagenesis" refers to an algorithm for protein engineering (protein mutagenesis) developed to produce diverse populations of phenotypically related mutants whose members differ in amino acid sequence. This method uses a feedback mechanism to control successive rounds of combinatorial cassette mutagenesis. *See, e.g.*, Arkin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 89: 7811-7815 (1992).

The term "exponential ensemble mutagenesis" refers to a process for generating combinatorial libraries with a high percentage of unique and functional mutants, wherein

small groups of residues are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. *See, e.g.*, Delegrave *et al.*, *Biotechnology Research* 11: 1548-1552 (1993); Arnold, *Current Opinion in Biotechnology* 4: 450-455 (1993). Each of the references mentioned above are hereby incorporated by reference in its entirety.

"Operatively linked" expression control sequences refers to a linkage in which the expression control sequence is contiguous with the gene of interest to control the gene of interest, as well as expression control sequences that act in *trans* or at a distance to control the gene of interest.

The term "expression control sequence" as used herein refers to polynucleotide sequences which are necessary to affect the expression of coding sequences to which they are operatively linked. Expression control sequences are sequences which control the transcription, post-transcriptional events and translation of nucleic acid sequences. Expression control sequences include appropriate transcription initiation, termination, promoter and enhancer sequences; efficient RNA processing signals such as splicing and polyadenylation signals; sequences that stabilize cytoplasmic mRNA; sequences that enhance translation efficiency (*e.g.*, ribosome binding sites); sequences that enhance protein stability; and when desired, sequences that enhance protein secretion. The nature of such control sequences differs depending upon the host organism; in prokaryotes, such control sequences generally include the promoter, ribosomal binding site, and transcription termination sequence. The term "control sequences" is intended to include, at a minimum, all components whose presence is essential for expression, and can also include additional components whose presence is advantageous, for example, leader sequences and fusion partner sequences.

The term "vector," as used herein, is intended to refer to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. One type of vector is a "plasmid", which refers to a circular double-stranded DNA loop into which additional DNA segments may be ligated. Other vectors include cosmids, bacterial artificial chromosomes (BAC) and yeast artificial chromosomes (YAC). Another type of vector is a viral vector, wherein additional DNA segments may be ligated into the viral genome. Viral vectors that infect bacterial cells are referred to as bacteriophages. Certain vectors are capable of autonomous replication in a host cell into which they are

introduced (*e.g.*, bacterial vectors having a bacterial origin of replication). Other vectors can be integrated into the genome of a host cell upon introduction into the host cell, and thereby are replicated along with the host genome. Moreover, certain vectors are capable of directing the expression of genes to which they are operatively linked. Such vectors are referred to herein as "recombinant expression vectors" (or simply, "expression vectors"). In general, expression vectors of utility in recombinant DNA techniques are often in the form of plasmids. In the present specification, "plasmid" and "vector" may be used interchangeably as the plasmid is the most commonly used form of vector. However, the invention is intended to include other forms of expression vectors that serve equivalent functions.

The term "recombinant host cell" (or simply "host cell"), as used herein, is intended to refer to a cell into which an expression vector has been introduced. It should be understood that such terms are intended to refer not only to the particular subject cell but to the progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term "host cell" as used herein.

As used herein, the phrase "open reading frame" and the equivalent acronym "ORF" refer to that portion of a transcript-derived nucleic acid that can be translated in its entirety into a sequence of contiguous amino acids. As so defined, an ORF has length, measured in nucleotides, exactly divisible by 3. As so defined, an ORF need not encode the entirety of a natural protein.

As used herein, the phrase "ORF-encoded peptide" refers to the predicted or actual translation of an ORF.

As used herein, the phrase "degenerate variant" of a reference nucleic acid sequence intends all nucleic acid sequences that can be directly translated, using the standard genetic code, to provide an amino acid sequence identical to that translated from the reference nucleic acid sequence.

The term "polypeptide" encompasses both naturally-occurring and non-naturally-occurring proteins and polypeptides, polypeptide fragments and polypeptide mutants, derivatives and analogs. A polypeptide may be monomeric or polymeric. Further, a polypeptide may comprise a number of different modules within a single polypeptide

each of which has one or more distinct activities. A preferred polypeptide in accordance with the invention comprises a CSP encoded by a nucleic acid molecule of the instant invention, as well as a fragment, mutant, analog and derivative thereof.

The term "isolated protein" or "isolated polypeptide" is a protein or polypeptide that by virtue of its origin or source of derivation (1) is not associated with naturally associated components that accompany it in its native state, (2) is free of other proteins from the same species (3) is expressed by a cell from a different species, or (4) does not occur in nature. Thus, a polypeptide that is chemically synthesized or synthesized in a cellular system different from the cell from which it naturally originates will be "isolated" from its naturally associated components. A polypeptide or protein may also be rendered substantially free of naturally associated components by isolation, using protein purification techniques well-known in the art.

A protein or polypeptide is "substantially pure," "substantially homogeneous" or "substantially purified" when at least about 60% to 75% of a sample exhibits a single species of polypeptide. The polypeptide or protein may be monomeric or multimeric. A substantially pure polypeptide or protein will typically comprise about 50%, 60%, 70%, 80% or 90% W/W of a protein sample, more usually about 95%, and preferably will be over 99% pure. Protein purity or homogeneity may be indicated by a number of means well-known in the art, such as polyacrylamide gel electrophoresis of a protein sample, followed by visualizing a single polypeptide band upon staining the gel with a stain well-known in the art. For certain purposes, higher resolution may be provided by using HPLC or other means well-known in the art for purification.

The term "polypeptide fragment" as used herein refers to a polypeptide of the instant invention that has an amino-terminal and/or carboxy-terminal deletion compared to a full-length polypeptide. In a preferred embodiment, the polypeptide fragment is a contiguous sequence in which the amino acid sequence of the fragment is identical to the corresponding positions in the naturally-occurring sequence. Fragments typically are at least 5, 6, 7, 8, 9 or 10 amino acids long, preferably at least 12, 14, 16 or 18 amino acids long, more preferably at least 20 amino acids long, more preferably at least 25, 30, 35, 40 or 45, amino acids, even more preferably at least 50 or 60 amino acids long, and even more preferably at least 70 amino acids long.

A "derivative" refers to polypeptides or fragments thereof that are substantially similar in primary structural sequence but which include, *e.g.*, *in vivo* or *in vitro* chemical and biochemical modifications that are not found in the native polypeptide. Such modifications include, for example, acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of phosphotidylinositol, cross-linking, cyclization, disulfide bond formation, demethylation, formation of covalent cross-links, formation of cystine, formation of pyroglutamate, formylation, gamma-carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristoylation, oxidation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination. Other modification include, *e.g.*, labeling with radionuclides, and various enzymatic modifications, as will be readily appreciated by those skilled in the art. A variety of methods for labeling polypeptides and of substituents or labels useful for such purposes are well-known in the art, and include radioactive isotopes such as ^{125}I, ^{32}P, ^{35}S, and ^{3}H, ligands which bind to labeled antiligands (*e.g.*, antibodies), fluorophores, chemiluminescent agents, enzymes, and antiligands which can serve as specific binding pair members for a labeled ligand. The choice of label depends on the sensitivity required, ease of conjugation with the primer, stability requirements, and available instrumentation. Methods for labeling polypeptides are well-known in the art. *See* Ausubel (1992), *supra*; Ausubel (1999), *supra*, herein incorporated by reference.

The term "fusion protein" refers to polypeptides of the instant invention comprising polypeptides or fragments coupled to heterologous amino acid sequences. Fusion proteins are useful because they can be constructed to contain two or more desired functional elements from two or more different proteins. A fusion protein comprises at least 10 contiguous amino acids from a polypeptide of interest, more preferably at least 20 or 30 amino acids, even more preferably at least 40, 50 or 60 amino acids, yet more preferably at least 75, 100 or 125 amino acids. Fusion proteins can be produced recombinantly by constructing a nucleic acid sequence which encodes the polypeptide or a fragment thereof in frame with a nucleic acid sequence encoding a

different protein or peptide and then expressing the fusion protein. Alternatively, a fusion protein can be produced chemically by crosslinking the polypeptide or a fragment thereof to another protein.

The term "analog" refers to both polypeptide analogs and non-peptide analogs. The term "polypeptide analog" as used herein refers to a polypeptide of the instant invention that is comprised of a segment of at least 25 amino acids that has substantial identity to a portion of an amino acid sequence but which contains non-natural amino acids or non-natural inter-residue bonds. In a preferred embodiment, the analog has the same or similar biological activity as the native polypeptide. Typically, polypeptide analogs comprise a conservative amino acid substitution (or insertion or deletion) with respect to the naturally-occurring sequence. Analogs typically are at least 20 amino acids long, preferably at least 50 amino acids long or longer, and can often be as long as a full-length naturally-occurring polypeptide.

The term "non-peptide analog" refers to a compound with properties that are analogous to those of a reference polypeptide of the instant invention. A non-peptide compound may also be termed a "peptide mimetic" or a "peptidomimetic." Such compounds are often developed with the aid of computerized molecular modeling. Peptide mimetics that are structurally similar to useful peptides may be used to produce an equivalent effect. Generally, peptidomimetics are structurally similar to a paradigm polypeptide (*i.e.*, a polypeptide that has a desired biochemical property or pharmacological activity), but have one or more peptide linkages optionally replaced by a linkage selected from the group consisting of: $--CH_2NH--$, $--CH_2S--$, $--CH_2-CH_2--$, $--CH=CH--$(cis and trans), $--COCH_2--$, $--CH(OH)CH_2--$, and $-CH_2SO--$, by methods well-known in the art. Systematic substitution of one or more amino acids of a consensus sequence with a D-amino acid of the same type (*e.g.*, D-lysine in place of L-lysine) may also be used to generate more stable peptides. In addition, constrained peptides comprising a consensus sequence or a substantially identical consensus sequence variation may be generated by methods known in the art (Rizo *et al.*, *Ann. Rev. Biochem.* 61:387-418 (1992), incorporated herein by reference). For example, one may add internal cysteine residues capable of forming intramolecular disulfide bridges which cyclize the peptide.

A "polypeptide mutant" or "mutein" refers to a polypeptide of the instant invention whose sequence contains substitutions, insertions or deletions of one or more amino acids compared to the amino acid sequence of a native or wild-type protein. A mutein may have one or more amino acid point substitutions, in which a single amino acid at a position has been changed to another amino acid, one or more insertions and/or deletions, in which one or more amino acids are inserted or deleted, respectively, in the sequence of the naturally-occurring protein, and/or truncations of the amino acid sequence at either or both the amino or carboxy termini. Further, a mutein may have the same or different biological activity as the naturally-occurring protein. For instance, a mutein may have an increased or decreased biological activity. A mutein has at least 50% sequence similarity to the wild type protein, preferred is 60% sequence similarity, more preferred is 70% sequence similarity. Even more preferred are muteins having 80%, 85% or 90% sequence similarity to the wild type protein. In an even more preferred embodiment, a mutein exhibits 95% sequence identity, even more preferably 97%, even more preferably 98% and even more preferably 99%. Sequence similarity may be measured by any common sequence analysis algorithm, such as Gap or Bestfit.

Preferred amino acid substitutions are those which: (1) reduce susceptibility to proteolysis, (2) reduce susceptibility to oxidation, (3) alter binding affinity for forming protein complexes, (4) alter binding affinity or enzymatic activity, and (5) confer or modify other physicochemical or functional properties of such analogs. For example, single or multiple amino acid substitutions (preferably conservative amino acid substitutions) may be made in the naturally-occurring sequence (preferably in the portion of the polypeptide outside the domain(s) forming intermolecular contacts. In a preferred embodiment, the amino acid substitutions are moderately conservative substitutions or conservative substitutions. In a more preferred embodiment, the amino acid substitutions are conservative substitutions. A conservative amino acid substitution should not substantially change the structural characteristics of the parent sequence (*e.g.*, a replacement amino acid should not tend to disrupt a helix that occurs in the parent sequence, or disrupt other types of secondary structure that characterizes the parent sequence). Examples of art-recognized polypeptide secondary and tertiary structures are described in Creighton (ed.), Proteins, Structures and Molecular Principles, W. H. Freeman and Company (1984); Branden *et al.* (ed.), Introduction to Protein Structure,

Garland Publishing (1991); Thornton *et al.*, *Nature* 354:105-106 (1991), each of which are incorporated herein by reference.

As used herein, the twenty conventional amino acids and their abbreviations follow conventional usage. *See* Golub *et al.* (eds.), Immunology - A Synthesis 2nd Ed., Sinauer Associates (1991), which is incorporated herein by reference. Stereoisomers (*e.g.*, D-amino acids) of the twenty conventional amino acids, unnatural amino acids such as -, -disubstituted amino acids, N-alkyl amino acids, and other unconventional amino acids may also be suitable components for polypeptides of the present invention. Examples of unconventional amino acids include: 4-hydroxyproline, γ-carboxyglutamate, -N,N,N-trimethyllysine, -N-acetyllysine, O-phosphoserine, N-acetylserine, N-formylmethionine, 3-methylhistidine, 5-hydroxylysine, s-N-methylarginine, and other similar amino acids and imino acids (*e.g.*, 4-hydroxyproline). In the polypeptide notation used herein, the lefthand direction is the amino terminal direction and the right hand direction is the carboxy-terminal direction, in accordance with standard usage and convention.

A protein has "homology" or is "homologous" to a protein from another organism if the encoded amino acid sequence of the protein has a similar sequence to the encoded amino acid sequence of a protein of a different organism and has a similar biological activity or function. Alternatively, a protein may have homology or be homologous to another protein if the two proteins have similar amino acid sequences and have similar biological activities or functions. Although two proteins are said to be "homologous," this does not imply that there is necessarily an evolutionary relationship between the proteins. Instead, the term "homologous" is defined to mean that the two proteins have similar amino acid sequences and similar biological activities or functions. In a preferred embodiment, a homologous protein is one that exhibits 50% sequence similarity to the wild type protein, preferred is 60% sequence similarity, more preferred is 70% sequence similarity. Even more preferred are homologous proteins that exhibit 80%, 85% or 90% sequence similarity to the wild type protein. In a yet more preferred embodiment, a homologous protein exhibits 95%, 97%, 98% or 99% sequence similarity.

When "sequence similarity" is used in reference to proteins or peptides, it is recognized that residue positions that are not identical often differ by conservative amino acid substitutions. In a preferred embodiment, a polypeptide that has "sequence

similarity" comprises conservative or moderately conservative amino acid substitutions. A "conservative amino acid substitution" is one in which an amino acid residue is substituted by another amino acid residue having a side chain (R group) with similar chemical properties (*e.g.*, charge or hydrophobicity). In general, a conservative amino acid substitution will not substantially change the functional properties of a protein. In cases where two or more amino acid sequences differ from each other by conservative substitutions, the percent sequence identity or degree of similarity may be adjusted upwards to correct for the conservative nature of the substitution. Means for making this adjustment are well-known to those of skill in the art. *See, e.g.*, Pearson, *Methods Mol. Biol.* 24: 307-31 (1994), herein incorporated by reference.

For instance, the following six groups each contain amino acids that are conservative substitutions for one another:

1) Serine (S), Threonine (T);
2) Aspartic Acid (D), Glutamic Acid (E);
3) Asparagine (N), Glutamine (Q);
4) Arginine (R), Lysine (K);
5) Isoleucine (I), Leucine (L), Methionine (M), Alanine (A), Valine (V), and
6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W).

Alternatively, a conservative replacement is any change having a positive value in the PAM250 log-likelihood matrix disclosed in Gonnet *et al.*, *Science* 256: 1443-45 (1992), herein incorporated by reference. A "moderately conservative" replacement is any change having a nonnegative value in the PAM250 log-likelihood matrix.

Sequence similarity for polypeptides, which is also referred to as sequence identity, is typically measured using sequence analysis software. Protein analysis software matches similar sequences using measures of similarity assigned to various substitutions, deletions and other modifications, including conservative amino acid substitutions. For instance, GCG contains programs such as "Gap" and "Bestfit" which can be used with default parameters to determine sequence homology or sequence identity between closely related polypeptides, such as homologous polypeptides from different species of organisms or between a wild type protein and a mutein thereof. *See, e.g.*, GCG Version 6.1. Other programs include FASTA, discussed *supra*.

A preferred algorithm when comparing a sequence of the invention to a database containing a large number of sequences from different organisms is the computer program BLAST, especially blastp or tblastn. *See, e.g.,* Altschul *et al., J. Mol. Biol.* 215: 403-410 (1990); Altschul *et al., Nucleic Acids Res.* 25:3389-402 (1997); herein incorporated by reference. Preferred parameters for blastp are:

Expectation value:	10 (default)
Filter:	seg (default)
Cost to open a gap:	11 (default)
Cost to extend a gap:	1 (default
Max. alignments:	100 (default)
Word size:	11 (default)
No. of descriptions:	100 (default)
Penalty Matrix:	BLOSUM62

The length of polypeptide sequences compared for homology will generally be at least about 16 amino acid residues, usually at least about 20 residues, more usually at least about 24 residues, typically at least about 28 residues, and preferably more than about 35 residues. When searching a database containing sequences from a large number of different organisms, it is preferable to compare amino acid sequences.

Database searching using amino acid sequences can be measured by algorithms other than blastp are known in the art. For instance, polypeptide sequences can be compared using FASTA, a program in GCG Version 6.1. FASTA (*e.g.*, FASTA2 and FASTA3) provides alignments and percent sequence identity of the regions of the best overlap between the query and search sequences (Pearson (1990), *supra*; Pearson (2000), *supra*. For example, percent sequence identity between amino acid sequences can be determined using FASTA with its default or recommended parameters (a word size of 2 and the PAM250 scoring matrix), as provided in GCG Version 6.1, herein incorporated by reference.

An "antibody" refers to an intact immunoglobulin, or to an antigen-binding portion thereof that competes with the intact antibody for specific binding to a molecular species, *e.g.*, a polypeptide of the instant invention. Antigen-binding portions may be produced by recombinant DNA techniques or by enzymatic or chemical cleavage of intact antibodies. Antigen-binding portions include, *inter alia*, Fab, Fab', $F(ab')_2$, Fv,

dAb, and complementarity determining region (CDR) fragments, single-chain antibodies (scFv), chimeric antibodies, diabodies and polypeptides that contain at least a portion of an immunoglobulin that is sufficient to confer specific antigen binding to the polypeptide. An Fab fragment is a monovalent fragment consisting of the VL, VH, CL and CH1 domains; an $F(ab')_2$ fragment is a bivalent fragment comprising two Fab fragments linked by a disulfide bridge at the hinge region; an Fd fragment consists of the VH and CH1 domains; an Fv fragment consists of the VL and VH domains of a single arm of an antibody; and a dAb fragment consists of a VH domain. *See, e.g.,* Ward *et al.*, *Nature* 341: 544-546 (1989).

By "bind specifically" and "specific binding" is here intended the ability of the antibody to bind to a first molecular species in preference to binding to other molecular species with which the antibody and first molecular species are admixed. An antibody is said specifically to "recognize" a first molecular species when it can bind specifically to that first molecular species.

A single-chain antibody (scFv) is an antibody in which a VL and VH region are paired to form a monovalent molecule via a synthetic linker that enables them to be made as a single protein chain. *See, e.g.,* Bird *et al.*, *Science* 242: 423-426 (1988); Huston *et al.*, *Proc. Natl. Acad. Sci. USA* 85: 5879-5883 (1988). Diabodies are bivalent, bispecific antibodies in which VH and VL domains are expressed on a single polypeptide chain, but using a linker that is too short to allow for pairing between the two domains on the same chain, thereby forcing the domains to pair with complementary domains of another chain and creating two antigen binding sites. *See e.g.*, Holliger *et al.*, *Proc. Natl. Acad. Sci. USA* 90: 6444-6448 (1993); Poljak *et al.*, *Structure* 2: 1121-1123 (1994). One or more CDRs may be incorporated into a molecule either covalently or noncovalently to make it an immunoadhesin. An immunoadhesin may incorporate the CDR(s) as part of a larger polypeptide chain, may covalently link the CDR(s) to another polypeptide chain, or may incorporate the CDR(s) noncovalently. The CDRs permit the immunoadhesin to specifically bind to a particular antigen of interest. A chimeric antibody is an antibody that contains one or more regions from one antibody and one or more regions from one or more other antibodies.

An antibody may have one or more binding sites. If there is more than one binding site, the binding sites may be identical to one another or may be different. For

instance, a naturally-occurring immunoglobulin has two identical binding sites, a single-chain antibody or Fab fragment has one binding site, while a "bispecific" or "bifunctional" antibody has two different binding sites.

An "isolated antibody" is an antibody that (1) is not associated with naturally-associated components, including other naturally-associated antibodies, that accompany it in its native state, (2) is free of other proteins from the same species, (3) is expressed by a cell from a different species, or (4) does not occur in nature. It is known that purified proteins, including purified antibodies, may be stabilized with non-naturally-associated components. The non-naturally-associated component may be a protein, such as albumin (*e.g.*, BSA) or a chemical such as polyethylene glycol (PEG).

A "neutralizing antibody" or "an inhibitory antibody" is an antibody that inhibits the activity of a polypeptide or blocks the binding of a polypeptide to a ligand that normally binds to it. An "activating antibody" is an antibody that increases the activity of a polypeptide.

The term "epitope" includes any protein determinant capable of specifically binding to an immunoglobulin or T-cell receptor. Epitopic determinants usually consist of chemically active surface groupings of molecules such as amino acids or sugar side chains and usually have specific three-dimensional structural characteristics, as well as specific charge characteristics. An antibody is said to specifically bind an antigen when the dissociation constant is less than1 μM, preferably less than100 nM and most preferably less than 10 nM.

The term "patient" as used herein includes human and veterinary subjects.

Throughout this specification and claims, the word "comprise," or variations such as "comprises" or "comprising," will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers.

The term "colon specific" refers to a nucleic acid molecule or polypeptide that is expressed predominantly in the colon as compared to other tissues in the body. In a preferred embodiment, a "colon specific" nucleic acid molecule or polypeptide is expressed at a level that is 5-fold higher than any other tissue in the body. In a more preferred embodiment, the "colon specific" nucleic acid molecule or polypeptide is expressed at a level that is 10-fold higher than any other tissue in the body, more

preferably at least 15-fold, 20-fold, 25-fold, 50-fold or 100-fold higher than any other tissue in the body. Nucleic acid molecule levels may be measured by nucleic acid hybridization, such as Northern blot hybridization, or quantitative PCR. Polypeptide levels may be measured by any method known to accurately quantitate protein levels, such as Western blot analysis.

Nucleic Acid Molecules, Regulatory Sequences, Vectors, Host Cells and Recombinant Methods of Making Polypeptides

Nucleic Acid Molecules

One aspect of the invention provides isolated nucleic acid molecules that are specific to the colon or to colon cells or tissue or that are derived from such nucleic acid molecules. These isolated colon specific nucleic acids (CSNAs) may comprise a cDNA, a genomic DNA, RNA, or a fragment of one of these nucleic acids, or may be a non-naturally-occurring nucleic acid molecule. In a preferred embodiment, the nucleic acid molecule encodes a polypeptide that is specific to colon, a colon-specific polypeptide (CSP). In a more preferred embodiment, the nucleic acid molecule encodes a polypeptide that comprises an amino acid sequence of SEQ ID NO: 101 through 176. In another highly preferred embodiment, the nucleic acid molecule comprises a nucleic acid sequence of SEQ ID NO: 1 through 100.

A CSNA may be derived from a human or from another animal. In a preferred embodiment, the CSNA is derived from a human or other mammal. In a more preferred embodiment, the CSNA is derived from a human or other primate. In an even more preferred embodiment, the CSNA is derived from a human.

By "nucleic acid molecule" for purposes of the present invention, it is also meant to be inclusive of nucleic acid sequences that selectively hybridize to a nucleic acid molecule encoding a CSNA or a complement thereof. The hybridizing nucleic acid molecule may or may not encode a polypeptide or may not encode a CSP. However, in a preferred embodiment, the hybridizing nucleic acid molecule encodes a CSP. In a more preferred embodiment, the invention provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule that encodes a polypeptide comprising an amino acid sequence of SEQ ID NO: 101 through 176. In an even more preferred embodiment, the invention provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule comprising the nucleic acid sequence of SEQ ID NO: 1 through 100.

In a preferred embodiment, the nucleic acid molecule selectively hybridizes to a nucleic acid molecule encoding a CSP under low stringency conditions. In a more preferred embodiment, the nucleic acid molecule selectively hybridizes to a nucleic acid molecule encoding a CSP under moderate stringency conditions. In a more preferred embodiment, the nucleic acid molecule selectively hybridizes to a nucleic acid molecule encoding a CSP under high stringency conditions. In an even more preferred embodiment, the nucleic acid molecule hybridizes under low, moderate or high stringency conditions to a nucleic acid molecule encoding a polypeptide comprising an amino acid sequence of SEQ ID NO: 101 through 176. In a yet more preferred embodiment, the nucleic acid molecule hybridizes under low, moderate or high stringency conditions to a nucleic acid molecule comprising a nucleic acid sequence selected from SEQ ID NO: 1 through 100. In a preferred embodiment of the invention, the hybridizing nucleic acid molecule may be used to express recombinantly a polypeptide of the invention.

By "nucleic acid molecule" as used herein it is also meant to be inclusive of sequences that exhibits substantial sequence similarity to a nucleic acid encoding a CSP or a complement of the encoding nucleic acid molecule. In a preferred embodiment, the nucleic acid molecule exhibits substantial sequence similarity to a nucleic acid molecule encoding human CSP. In a more preferred embodiment, the nucleic acid molecule exhibits substantial sequence similarity to a nucleic acid molecule encoding a polypeptide having an amino acid sequence of SEQ ID NO: 101 through 176. In a preferred embodiment, the similar nucleic acid molecule is one that has at least 60% sequence identity with a nucleic acid molecule encoding a CSP, such as a polypeptide having an amino acid sequence of SEQ ID NO: 101 through 176, more preferably at least 70%, even more preferably at least 80% and even more preferably at least 85%. In a more preferred embodiment, the similar nucleic acid molecule is one that has at least 90% sequence identity with a nucleic acid molecule encoding a CSP, more preferably at least 95%, more preferably at least 97%, even more preferably at least 98%, and still more preferably at least 99%. In another highly preferred embodiment, the nucleic acid molecule is one that has at least 99.5%, 99.6%, 99.7%, 99.8% or 99.9% sequence identity with a nucleic acid molecule encoding a CSP.

In another preferred embodiment, the nucleic acid molecule exhibits substantial sequence similarity to a CSNA or its complement. In a more preferred embodiment, the nucleic acid molecule exhibits substantial sequence similarity to a nucleic acid molecule comprising a nucleic acid sequence of SEQ ID NO: 1 through 100. In a preferred embodiment, the nucleic acid molecule is one that has at least 60% sequence identity with a CSNA, such as one having a nucleic acid sequence of SEQ ID NO: 1 through 100, more preferably at least 70%, even more preferably at least 80% and even more preferably at least 85%. In a more preferred embodiment, the nucleic acid molecule is one that has at least 90% sequence identity with a CSNA, more preferably at least 95%, more preferably at least 97%, even more preferably at least 98%, and still more preferably at least 99%. In another highly preferred embodiment, the nucleic acid molecule is one that has at least 99.5%, 99.6%, 99.7%, 99.8% or 99.9% sequence identity with a CSNA.

A nucleic acid molecule that exhibits substantial sequence similarity may be one that exhibits sequence identity over its entire length to a CSNA or to a nucleic acid molecule encoding a CSP, or may be one that is similar over only a part of its length. In this case, the part is at least 50 nucleotides of the CSNA or the nucleic acid molecule encoding a CSP, preferably at least 100 nucleotides, more preferably at least 150 or 200 nucleotides, even more preferably at least 250 or 300 nucleotides, still more preferably at least 400 or 500 nucleotides.

The substantially similar nucleic acid molecule may be a naturally-occurring one that is derived from another species, especially one derived from another primate, wherein the similar nucleic acid molecule encodes an amino acid sequence that exhibits significant sequence identity to that of SEQ ID NO: 101 through 176 or demonstrates significant sequence identity to the nucleotide sequence of SEQ ID NO: 1 through 100. The similar nucleic acid molecule may also be a naturally-occurring nucleic acid molecule from a human, when the CSNA is a member of a gene family. The similar nucleic acid molecule may also be a naturally-occurring nucleic acid molecule derived from a non-primate, mammalian species, including without limitation, domesticated species, *e.g.*, dog, cat, mouse, rat, rabbit, hamster, cow, horse and pig; and wild animals, *e.g.*, monkey, fox, lions, tigers, bears, giraffes, zebras, etc. The substantially similar nucleic acid molecule may also be a naturally-occurring nucleic acid molecule derived

from a non-mammalian species, such as birds or reptiles. The naturally-occurring substantially similar nucleic acid molecule may be isolated directly from humans or other species. In another embodiment, the substantially similar nucleic acid molecule may be one that is experimentally produced by random mutation of a nucleic acid molecule. In another embodiment, the substantially similar nucleic acid molecule may be one that is experimentally produced by directed mutation of a CSNA. Further, the substantially similar nucleic acid molecule may or may not be a CSNA. However, in a preferred embodiment, the substantially similar nucleic acid molecule is a CSNA.

By "nucleic acid molecule" it is also meant to be inclusive of allelic variants of a CSNA or a nucleic acid encoding a CSP. For instance, single nucleotide polymorphisms (SNPs) occur frequently in eukaryotic genomes. In fact, more than 1.4 million SNPs have already identified in the human genome, International Human Genome Sequencing Consortium, *Nature* 409: 860-921 (2001). Thus, the sequence determined from one individual of a species may differ from other allelic forms present within the population. Additionally, small deletions and insertions, rather than single nucleotide polymorphisms, are not uncommon in the general population, and often do not alter the function of the protein. Further, amino acid substitutions occur frequently among natural allelic variants, and often do not substantially change protein function.

In a preferred embodiment, the nucleic acid molecule comprising an allelic variant is a variant of a gene, wherein the gene is transcribed into an mRNA that encodes a CSP. In a more preferred embodiment, the gene is transcribed into an mRNA that encodes a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176. In another preferred embodiment, the allelic variant is a variant of a gene, wherein the gene is transcribed into an mRNA that is a CSNA. In a more preferred embodiment, the gene is transcribed into an mRNA that comprises the nucleic acid sequence of SEQ ID NO: 1 through 100. In a preferred embodiment, the allelic variant is a naturally-occurring allelic variant in the species of interest. In a more preferred embodiment, the species of interest is human.

By "nucleic acid molecule" it is also meant to be inclusive of a part of a nucleic acid sequence of the instant invention. The part may or may not encode a polypeptide, and may or may not encode a polypeptide that is a CSP. However, in a preferred embodiment, the part encodes a CSP. In one aspect, the invention comprises a part of a

CSNA. In a second aspect, the invention comprises a part of a nucleic acid molecule that hybridizes or exhibits substantial sequence similarity to a CSNA. In a third aspect, the invention comprises a part of a nucleic acid molecule that is an allelic variant of a CSNA. In a fourth aspect, the invention comprises a part of a nucleic acid molecule that encodes a CSP. A part comprises at least 10 nucleotides, more preferably at least 15, 17, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400 or 500 nucleotides. The maximum size of a nucleic acid part is one nucleotide shorter than the sequence of the nucleic acid molecule encoding the full-length protein.

By "nucleic acid molecule" it is also meant to be inclusive of sequence that encoding a fusion protein, a homologous protein, a polypeptide fragment, a mutein or a polypeptide analog, as described below.

Nucleotide sequences of the instantly-described nucleic acids were determined by sequencing a DNA molecule that had resulted, directly or indirectly, from at least one enzymatic polymerization reaction (*e.g.*, reverse transcription and/or polymerase chain reaction) using an automated sequencer (such as the MegaBACE™ 1000, Molecular Dynamics, Sunnyvale, CA, USA). Further, all amino acid sequences of the polypeptides of the present invention were predicted by translation from the nucleic acid sequences so determined, unless otherwise specified.

In a preferred embodiment of the invention, the nucleic acid molecule contains modifications of the native nucleic acid molecule. These modifications include nonnative internucleoside bonds, post-synthetic modifications or altered nucleotide analogues. One having ordinary skill in the art would recognize that the type of modification that can be made will depend upon the intended use of the nucleic acid molecule. For instance, when the nucleic acid molecule is used as a hybridization probe, the range of such modifications will be limited to those that permit sequence-discriminating base pairing of the resulting nucleic acid. When used to direct expression of RNA or protein *in vitro* or *in vivo*, the range of such modifications will be limited to those that permit the nucleic acid to function properly as a polymerization substrate. When the isolated nucleic acid is used as a therapeutic agent, the modifications will be limited to those that do not confer toxicity upon the isolated nucleic acid.

In a preferred embodiment, isolated nucleic acid molecules can include nucleotide analogues that incorporate labels that are directly detectable, such as radiolabels or

fluorophores, or nucleotide analogues that incorporate labels that can be visualized in a subsequent reaction, such as biotin or various haptens. In a more preferred embodiment, the labeled nucleic acid molecule may be used as a hybridization probe.

Common radiolabeled analogues include those labeled with ^{33}P, ^{32}P, and ^{35}S, such as -^{32}P-dATP, -^{32}P-dCTP, -^{32}P-dGTP, -^{32}P-dTTP, -^{32}P-3'dATP, -^{32}P-ATP, -^{32}P-CTP, -^{32}P-GTP, -^{32}P-UTP, -^{35}S-dATP, α-^{35}S-GTP, α-^{33}P-dATP, and the like.

Commercially available fluorescent nucleotide analogues readily incorporated into the nucleic acids of the present invention include Cy3-dCTP, Cy3-dUTP, Cy5-dCTP, Cy3-dUTP (Amersham Pharmacia Biotech, Piscataway, New Jersey, USA), fluorescein-12-dUTP, tetramethylrhodamine-6-dUTP, Texas Red®-5-dUTP, Cascade Blue®-7-dUTP, BODIPY® FL-14-dUTP, BODIPY® TMR-14-dUTP, BODIPY® TR-14-dUTP, Rhodamine Green™-5-dUTP, Oregon Green® 488-5-dUTP, Texas Red®-12-dUTP, BODIPY® 630/650-14-dUTP, BODIPY® 650/665-14-dUTP, Alexa Fluor® 488-5-dUTP, Alexa Fluor® 532-5-dUTP, Alexa Fluor® 568-5-dUTP, Alexa Fluor® 594-5-dUTP, Alexa Fluor® 546-14-dUTP, fluorescein-12-UTP, tetramethylrhodamine-6-UTP, Texas Red®-5-UTP, Cascade Blue®-7-UTP, BODIPY® FL-14-UTP, BODIPY® TMR-14-UTP, BODIPY® TR-14-UTP, Rhodamine Green™-5-UTP, Alexa Fluor® 488-5-UTP, Alexa Fluor® 546-14-UTP (Molecular Probes, Inc. Eugene, OR, USA). One may also custom synthesize nucleotides having other fluorophores. *See* Henegariu *et al.*, *Nature Biotechnol.* 18: 345-348 (2000), the disclosure of which is incorporated herein by reference in its entirety.

Haptens that are commonly conjugated to nucleotides for subsequent labeling include biotin (biotin-11-dUTP, Molecular Probes, Inc., Eugene, OR, USA; biotin-21-UTP, biotin-21-dUTP, Clontech Laboratories, Inc., Palo Alto, CA, USA), digoxigenin (DIG-11-dUTP, alkali labile, DIG-11-UTP, Roche Diagnostics Corp., Indianapolis, IN, USA), and dinitrophenyl (dinitrophenyl-11-dUTP, Molecular Probes, Inc., Eugene, OR, USA).

Nucleic acid molecules can be labeled by incorporation of labeled nucleotide analogues into the nucleic acid. Such analogues can be incorporated by enzymatic polymerization, such as by nick translation, random priming, polymerase chain reaction (PCR), terminal transferase tailing, and end-filling of overhangs, for DNA molecules, and *in vitro* transcription driven, *e.g.*, from phage promoters, such as T7, T3, and SP6, for

RNA molecules. Commercial kits are readily available for each such labeling approach. Analogues can also be incorporated during automated solid phase chemical synthesis. Labels can also be incorporated after nucleic acid synthesis, with the 5' phosphate and 3' hydroxyl providing convenient sites for post-synthetic covalent attachment of detectable labels.

Other post-synthetic approaches also permit internal labeling of nucleic acids. For example, fluorophores can be attached using a cisplatin reagent that reacts with the N7 of guanine residues (and, to a lesser extent, adenine bases) in DNA, RNA, and PNA to provide a stable coordination complex between the nucleic acid and fluorophore label (Universal Linkage System) (available from Molecular Probes, Inc., Eugene, OR, USA and Amersham Pharmacia Biotech, Piscataway, NJ, USA); *see* Alers *et al.*, *Genes, Chromosomes & Cancer* 25: 301- 305 (1999); Jelsma *et al.*, *J. NIH Res.* 5: 82 (1994); Van Belkum *et al.*, *BioTechniques* 16: 148-153 (1994), incorporated herein by reference. As another example, nucleic acids can be labeled using a disulfide-containing linker (FastTag™ Reagent, Vector Laboratories, Inc., Burlingame, CA, USA) that is photo- or thermally-coupled to the target nucleic acid using aryl azide chemistry; after reduction, a free thiol is available for coupling to a hapten, fluorophore, sugar, affinity ligand, or other marker.

One or more independent or interacting labels can be incorporated into the nucleic acid molecules of the present invention. For example, both a fluorophore and a moiety that in proximity thereto acts to quench fluorescence can be included to report specific hybridization through release of fluorescence quenching or to report exonucleotidic excision. *See, e.g.,* Tyagi *et al.*, *Nature Biotechnol.* 14: 303-308 (1996); Tyagi *et al.*, *Nature Biotechnol.* 16: 49-53 (1998); Sokol *et al.*, *Proc. Natl. Acad. Sci. USA* 95: 11538-11543 (1998); Kostrikis *et al.*, *Science* 279: 1228-1229 (1998); Marras *et al.*, *Genet. Anal.* 14: 151-156 (1999); U. S. Patent 5,846,726; 5,925,517; 5,925,517; 5,723,591 and 5,538,848; Holland *et al.*, *Proc. Natl. Acad. Sci. USA* 88: 7276-7280 (1991); Heid *et al.*, *Genome Res.* 6(10): 986-94 (1996); Kuimelis *et al.*, *Nucleic Acids Symp. Ser.* (37): 255-6 (1997); the disclosures of which are incorporated herein by reference in their entireties.

Nucleic acid molecules of the invention may be modified by altering one or more native phosphodiester internucleoside bonds to more nuclease-resistant, internucleoside

bonds. *See* Hartmann *et al.* (eds.), Manual of Antisense Methodology: Perspectives in Antisense Science, Kluwer Law International (1999); Stein *et al.* (eds.), Applied Antisense Oligonucleotide Technology, Wiley-Liss (1998); Chadwick *et al.* (eds.), Oligonucleotides as Therapeutic Agents - Symposium No. 209, John Wiley & Son Ltd (1997); the disclosures of which are incorporated herein by reference in their entireties. Such altered internucleoside bonds are often desired for antisense techniques or for targeted gene correction. *See* Gamper *et al., Nucl. Acids Res.* 28(21): 4332-4339 (2000), the disclosure of which is incorporated herein by reference in its entirety.

Modified oligonucleotide backbones include, without limitation, phosphorothioates, chiral phosphorothioates, phosphorodithioates, phosphotriesters, aminoalkylphosphotriesters, methyl and other alkyl phosphonates including 3'-alkylene phosphonates and chiral phosphonates, phosphinates, phosphoramidates including 3'-amino phosphoramidate and aminoalkylphosphoramidates, thionophosphoramidates, thionoalkylphosphonates, thionoalkylphosphotriesters, and boranophosphates having normal 3'-5' linkages, 2'-5' linked analogs of these, and those having inverted polarity wherein the adjacent pairs of nucleoside units are linked 3'-5' to 5'-3' or 2'-5' to 5'-2'. Representative United States patents that teach the preparation of the above phosphorus-containing linkages include, but are not limited to, U. S. Patents 3,687,808; 4,469,863; 4,476,301; 5,023,243; 5,177,196; 5,188,897; 5,264,423; 5,276,019; 5,278,302; 5,286,717; 5,321,131; 5,399,676; 5,405,939; 5,453,496; 5,455,233; 5,466,677; 5,476,925; 5,519,126; 5,536,821; 5,541,306; 5,550,111; 5,563,253; 5,571,799; 5,587,361; and 5,625,050, the disclosures of which are incorporated herein by reference in their entireties. In a preferred embodiment, the modified internucleoside linkages may be used for antisense techniques.

Other modified oligonucleotide backbones do not include a phosphorus atom, but have backbones that are formed by short chain alkyl or cycloalkyl internucleoside linkages, mixed heteroatom and alkyl or cycloalkyl internucleoside linkages, or one or more short chain heteroatomic or heterocyclic internucleoside linkages. These include those having morpholino linkages (formed in part from the sugar portion of a nucleoside); siloxane backbones; sulfide, sulfoxide and sulfone backbones; formacetyl and thioformacetyl backbones; methylene formacetyl and thioformacetyl backbones; alkene containing backbones; sulfamate backbones; methyleneimino and

methylenehydrazino backbones; sulfonate and sulfonamide backbones; amide backbones; and others having mixed N, O, S and CH_2 component parts. Representative U.S. patents that teach the preparation of the above backbones include, but are not limited to, U.S. Patent 5,034,506; 5,166,315; 5,185,444; 5,214,134; 5,216,141; 5,235,033; 5,264,562; 5,264,564; 5,405,938; 5,434,257; 5,466,677; 5,470,967; 5,489,677; 5,541,307; 5,561,225; 5,596,086; 5,602,240; 5,610,289; 5,602,240; 5,608,046; 5,610,289; 5,618,704; 5,623,070; 5,663,312; 5,633,360; 5,677,437 and 5,677,439; the disclosures of which are incorporated herein by reference in their entireties.

In other preferred oligonucleotide mimetics, both the sugar and the internucleoside linkage are replaced with novel groups, such as peptide nucleic acids (PNA). In PNA compounds, the phosphodiester backbone of the nucleic acid is replaced with an amide-containing backbone, in particular by repeating N-(2-aminoethyl) glycine units linked by amide bonds. Nucleobases are bound directly or indirectly to aza nitrogen atoms of the amide portion of the backbone, typically by methylene carbonyl linkages. PNA can be synthesized using a modified peptide synthesis protocol. PNA oligomers can be synthesized by both Fmoc and tBoc methods. Representative U.S. patents that teach the preparation of PNA compounds include, but are not limited to, U.S Patent 5,539,082; 5,714,331; and 5,719,262, each of which is herein incorporated by reference. Automated PNA synthesis is readily achievable on commercial synthesizers (*see, e.g.*, "PNA User's Guide," Rev. 2, February 1998, Perseptive Biosystems Part No. 60138, Applied Biosystems, Inc., Foster City, CA).

PNA molecules are advantageous for a number of reasons. First, because the PNA backbone is uncharged, PNA/DNA and PNA/RNA duplexes have a higher thermal stability than is found in DNA/DNA and DNA/RNA duplexes. The Tm of a PNA/DNA or PNA/RNA duplex is generally 1°C higher per base pair than the Tm of the corresponding DNA/DNA or DNA/RNA duplex (in 100 mM NaCl). Second, PNA molecules can also form stable PNA/DNA complexes at low ionic strength, under conditions in which DNA/DNA duplex formation does not occur. Third, PNA also demonstrates greater specificity in binding to complementary DNA because a PNA/DNA mismatch is more destabilizing than DNA/DNA mismatch. A single mismatch in mixed a PNA/DNA 15-mer lowers the Tm by 8–20°C (15°C on average). In the corresponding DNA/DNA duplexes, a single mismatch lowers the Tm by 4–16°C (11°C on average).

Because PNA probes can be significantly shorter than DNA probes, their specificity is greater. Fourth, PNA oligomers are resistant to degradation by enzymes, and the lifetime of these compounds is extended both *in vivo* and *in vitro* because nucleases and proteases do not recognize the PNA polyamide backbone with nucleobase sidechains. *See, e.g.*, Ray *et al.*, *FASEB J.* 14(9): 1041-60 (2000); Nielsen *et al.*, *Pharmacol Toxicol.* 86(1): 3-7 (2000); Larsen *et al.*, *Biochim Biophys Acta.* 1489(1): 159-66 (1999); Nielsen, *Curr. Opin. Struct. Biol.* 9(3): 353-7 (1999), and Nielsen, *Curr. Opin. Biotechnol.* 10(1): 71-5 (1999), the disclosures of which are incorporated herein by reference in their entireties.

Nucleic acid molecules may be modified compared to their native structure throughout the length of the nucleic acid molecule or can be localized to discrete portions thereof. As an example of the latter, chimeric nucleic acids can be synthesized that have discrete DNA and RNA domains and that can be used for targeted gene repair and modified PCR reactions, as further described in U.S. Patents 5,760,012 and 5,731,181, Misra *et al.*, *Biochem.* 37: 1917-1925 (1998); and Finn *et al.*, *Nucl. Acids Res.* 24: 3357-3363 (1996), the disclosures of which are incorporated herein by reference in their entireties.

Unless otherwise specified, nucleic acids of the present invention can include any topological conformation appropriate to the desired use; the term thus explicitly comprehends, among others, single-stranded, double-stranded, triplexed, quadruplexed, partially double-stranded, partially-triplexed, partially-quadruplexed, branched, hairpinned, circular, and padlocked conformations. Padlock conformations and their utilities are further described in Banér *et al.*, *Curr. Opin. Biotechnol.* 12: 11-15 (2001); Escude *et al.*, *Proc. Natl. Acad. Sci. USA* 14: 96(19):10603-7 (1999); Nilsson *et al.*, *Science* 265(5181): 2085-8 (1994), the disclosures of which are incorporated herein by reference in their entireties. Triplex and quadruplex conformations, and their utilities, are reviewed in Praseuth *et al.*, *Biochim. Biophys. Acta.* 1489(1): 181-206 (1999); Fox, *Curr. Med. Chem.* 7(1): 17-37 (2000); Kochetkova *et al.*, *Methods Mol. Biol.* 130: 189-201 (2000); Chan *et al.*, *J. Mol. Med.* 75(4): 267-82 (1997), the disclosures of which are incorporated herein by reference in their entireties.

Methods for Using Nucleic Acid Molecules as Probes and Primers

The isolated nucleic acid molecules of the present invention can be used as hybridization probes to detect, characterize, and quantify hybridizing nucleic acids in, and isolate hybridizing nucleic acids from, both genomic and transcript-derived nucleic acid samples. When free in solution, such probes are typically, but not invariably, detectably labeled; bound to a substrate, as in a microarray, such probes are typically, but not invariably unlabeled.

In one embodiment, the isolated nucleic acids of the present invention can be used as probes to detect and characterize gross alterations in the gene of a CSNA, such as deletions, insertions, translocations, and duplications of the CSNA genomic locus through fluorescence *in situ* hybridization (FISH) to chromosome spreads. *See, e.g.*, Andreeff *et al.* (eds.), Introduction to Fluorescence *In Situ* Hybridization: Principles and Clinical Applications, John Wiley & Sons (1999), the disclosure of which is incorporated herein by reference in its entirety. The isolated nucleic acids of the present invention can be used as probes to assess smaller genomic alterations using, *e.g.*, Southern blot detection of restriction fragment length polymorphisms. The isolated nucleic acid molecules of the present invention can be used as probes to isolate genomic clones that include the nucleic acid molecules of the present invention, which thereafter can be restriction mapped and sequenced to identify deletions, insertions, translocations, and substitutions (single nucleotide polymorphisms, SNPs) at the sequence level.

In another embodiment, the isolated nucleic acid molecules of the present invention can be used as probes to detect, characterize, and quantify CSNA in, and isolate CSNA from, transcript-derived nucleic acid samples. In one aspect, the isolated nucleic acid molecules of the present invention can be used as hybridization probes to detect, characterize by length, and quantify mRNA by Northern blot of total or poly-A^+-selected RNA samples. In another aspect, the isolated nucleic acid molecules of the present invention can be used as hybridization probes to detect, characterize by location, and quantify mRNA by *in situ* hybridization to tissue sections. *See, e.g.*, Schwarchzacher *et al.*, In Situ Hybridization, Springer-Verlag New York (2000), the disclosure of which is incorporated herein by reference in its entirety. In another preferred embodiment, the isolated nucleic acid molecules of the present invention can be used as hybridization probes to measure the representation of clones in a cDNA library or to isolate hybridizing

nucleic acid molecules acids from cDNA libraries, permitting sequence level characterization of mRNAs that hybridize to CSNAs, including, without limitations, identification of deletions, insertions, substitutions, truncations, alternatively spliced forms and single nucleotide polymorphisms. In yet another preferred embodiment, the nucleic acid molecules of the instant invention may be used in microarrays.

All of the aforementioned probe techniques are well within the skill in the art, and are described at greater length in standard texts such as Sambrook (2001), *supra*; Ausubel (1999), *supra*; and Walker *et al.* (eds.), The Nucleic Acids Protocols Handbook, Humana Press (2000), the disclosures of which are incorporated herein by reference in their entirety.

Thus, in one embodiment, a nucleic acid molecule of the invention may be used as a probe or primer to identify or amplify a second nucleic acid molecule that selectively hybridizes to the nucleic acid molecule of the invention. In a preferred embodiment, the probe or primer is derived from a nucleic acid molecule encoding a CSP. In a more preferred embodiment, the probe or primer is derived from a nucleic acid molecule encoding a polypeptide having an amino acid sequence of SEQ ID NO: 101 through 176. In another preferred embodiment, the probe or primer is derived from a CSNA. In a more preferred embodiment, the probe or primer is derived from a nucleic acid molecule having a nucleotide sequence of SEQ ID NO: 1 through 100.

In general, a probe or primer is at least 10 nucleotides in length, more preferably at least 12, more preferably at least 14 and even more preferably at least 16 or 17 nucleotides in length. In an even more preferred embodiment, the probe or primer is at least 18 nucleotides in length, even more preferably at least 20 nucleotides and even more preferably at least 22 nucleotides in length. Primers and probes may also be longer in length. For instance, a probe or primer may be 25 nucleotides in length, or may be 30, 40 or 50 nucleotides in length. Methods of performing nucleic acid hybridization using oligonucleotide probes are well-known in the art. *See, e.g.*, Sambrook *et al.*, 1989, *supra*, Chapter 11 and pp. 11.31-11.32 and 11.40-11.44, which describes radiolabeling of short probes, and pp. 11.45-11.53, which describe hybridization conditions for oligonucleotide probes, including specific conditions for probe hybridization (pp. 11.50-11.51).

Methods of performing primer-directed amplification are also well-known in the art. Methods for performing the polymerase chain reaction (PCR) are compiled, *inter*

alia, in McPherson, PCR Basics: From Background to Bench, Springer Verlag (2000); Innis *et al.* (eds.), PCR Applications: Protocols for Functional Genomics, Academic Press (1999); Gelfand *et al.* (eds.), PCR Strategies, Academic Press (1998); Newton *et al.*, PCR, Springer-Verlag New York (1997); Burke (ed.), PCR: Essential Techniques, John Wiley & Son Ltd (1996); White (ed.), PCR Cloning Protocols: From Molecular Cloning to Genetic Engineering, Vol. 67, Humana Press (1996); McPherson *et al.* (eds.), PCR 2: A Practical Approach, Oxford University Press, Inc. (1995); the disclosures of which are incorporated herein by reference in their entireties. Methods for performing RT-PCR are collected, *e.g.*, in Siebert *et al.* (eds.), Gene Cloning and Analysis by RT-PCR, Eaton Publishing Company/Bio Techniques Books Division, 1998; Siebert (ed.), PCR Technique:RT-PCR, Eaton Publishing Company/ BioTechniques Books (1995); the disclosure of which is incorporated herein by reference in its entirety.

PCR and hybridization methods may be used to identify and/or isolate allelic variants, homologous nucleic acid molecules and fragments of the nucleic acid molecules of the invention. PCR and hybridization methods may also be used to identify, amplify and/or isolate nucleic acid molecules that encode homologous proteins, analogs, fusion protein or muteins of the invention. The nucleic acid primers of the present invention can be used to prime amplification of nucleic acid molecules of the invention, using transcript-derived or genomic DNA as template.

The nucleic acid primers of the present invention can also be used, for example, to prime single base extension (SBE) for SNP detection (*See, e.g.,* U.S. Patent 6,004,744, the disclosure of which is incorporated herein by reference in its entirety).

Isothermal amplification approaches, such as rolling circle amplification, are also now well-described. *See, e.g.,* Schweitzer *et al., Curr. Opin. Biotechnol.* 12(1): 21-7 (2001); U.S. Patents 5,854,033 and 5,714,320; and international patent publications WO 97/19193 and WO 00/15779, the disclosures of which are incorporated herein by reference in their entireties. Rolling circle amplification can be combined with other techniques to facilitate SNP detection. *See, e.g.,* Lizardi *et al., Nature Genet.* 19(3): 225-32 (1998).

Nucleic acid molecules of the present invention may be bound to a substrate either covalently or noncovalently. The substrate can be porous or solid, planar or non-planar, unitary or distributed. The bound nucleic acid molecules may be used as

hybridization probes, and may be labeled or unlabeled. In a preferred embodiment, the bound nucleic acid molecules are unlabeled.

In one embodiment, the nucleic acid molecule of the present invention is bound to a porous substrate, *e.g.*, a membrane, typically comprising nitrocellulose, nylon, or positively-charged derivatized nylon. The nucleic acid molecule of the present invention can be used to detect a hybridizing nucleic acid molecule that is present within a labeled nucleic acid sample, *e.g.*, a sample of transcript-derived nucleic acids. In another embodiment, the nucleic acid molecule is bound to a solid substrate, including, without limitation, glass, amorphous silicon, crystalline silicon or plastics. Examples of plastics include, without limitation, polymethylacrylic, polyethylene, polypropylene, polyacrylate, polymethylmethacrylate, polyvinylchloride, polytetrafluoroethylene, polystyrene, polycarbonate, polyacetal, polysulfone, celluloseacetate, cellulosenitrate, nitrocellulose, or mixtures thereof. The solid substrate may be any shape, including rectangular, disk-like and spherical. In a preferred embodiment, the solid substrate is a microscope slide or slide-shaped substrate.

The nucleic acid molecule of the present invention can be attached covalently to a surface of the support substrate or applied to a derivatized surface in a chaotropic agent that facilitates denaturation and adherence by presumed noncovalent interactions, or some combination thereof. The nucleic acid molecule of the present invention can be bound to a substrate to which a plurality of other nucleic acids are concurrently bound, hybridization to each of the plurality of bound nucleic acids being separately detectable. At low density, *e.g.* on a porous membrane, these substrate-bound collections are typically denominated macroarrays; at higher density, typically on a solid support, such as glass, these substrate bound collections of plural nucleic acids are colloquially termed microarrays. As used herein, the term microarray includes arrays of all densities. It is, therefore, another aspect of the invention to provide microarrays that include the nucleic acids of the present invention.

Expression Vectors, Host Cells and Recombinant Methods of Producing Polypeptides

Another aspect of the present invention relates to vectors that comprise one or more of the isolated nucleic acid molecules of the present invention, and host cells in which such vectors have been introduced.

The vectors can be used, *inter alia,* for propagating the nucleic acids of the present invention in host cells (cloning vectors), for shuttling the nucleic acids of the present invention between host cells derived from disparate organisms (shuttle vectors), for inserting the nucleic acids of the present invention into host cell chromosomes (insertion vectors), for expressing sense or antisense RNA transcripts of the nucleic acids of the present invention *in vitro* or within a host cell, and for expressing polypeptides encoded by the nucleic acids of the present invention, alone or as fusions to heterologous polypeptides (expression vectors). Vectors of the present invention will often be suitable for several such uses.

Vectors are by now well-known in the art, and are described, *inter alia*, in Jones *et al.* (eds.), Vectors: Cloning Applications: Essential Techniques (Essential Techniques Series), John Wiley & Son Ltd. (1998); Jones *et al.* (eds.), Vectors: Expression Systems: Essential Techniques (Essential Techniques Series), John Wiley & Son Ltd. (1998); Gacesa *et al.*, Vectors: Essential Data, John Wiley & Sons Ltd. (1995); Cid-Arregui (eds.), Viral Vectors: Basic Science and Gene Therapy, Eaton Publishing Co. (2000); Sambrook (2001), *supra*; Ausubel (1999), *supra*; the disclosures of which are incorporated herein by reference in their entireties. Furthermore, an enormous variety of vectors are available commercially. Use of existing vectors and modifications thereof being well within the skill in the art, only basic features need be described here.

Nucleic acid sequences may be expressed by operatively linking them to an expression control sequence in an appropriate expression vector and employing that expression vector to transform an appropriate unicellular host. Expression control sequences are sequences which control the transcription, post-transcriptional events and translation of nucleic acid sequences. Such operative linking of a nucleic sequence of this invention to an expression control sequence, of course, includes, if not already part of the nucleic acid sequence, the provision of a translation initiation codon, ATG or GTG, in the correct reading frame upstream of the nucleic acid sequence.

A wide variety of host/expression vector combinations may be employed in expressing the nucleic acid sequences of this invention. Useful expression vectors, for example, may consist of segments of chromosomal, non-chromosomal and synthetic nucleic acid sequences.

In one embodiment, prokaryotic cells may be used with an appropriate vector. Prokaryotic host cells are often used for cloning and expression. In a preferred embodiment, prokaryotic host cells include *E. coli*, *Pseudomonas*, *Bacillus* and *Streptomyces*. In a preferred embodiment, bacterial host cells are used to express the nucleic acid molecules of the instant invention. Useful expression vectors for bacterial hosts include bacterial plasmids, such as those from *E. coli*, *Bacillus* or *Streptomyces*, including pBluescript, pGEX-2T, pUC vectors, col E1, pCR1, pBR322, pMB9 and their derivatives, wider host range plasmids, such as RP4, phage DNAs, *e.g.*, the numerous derivatives of phage lambda, *e.g.*, NM989, λGT10 and λGT11, and other phages, *e.g.*, M13 and filamentous single-stranded phage DNA. Where *E. coli* is used as host, selectable markers are, analogously, chosen for selectivity in gram negative bacteria: *e.g.*, typical markers confer resistance to antibiotics, such as ampicillin, tetracycline, chloramphenicol, kanamycin, streptomycin and zeocin; auxotrophic markers can also be used.

In other embodiments, eukaryotic host cells, such as yeast, insect, mammalian or plant cells, may be used. Yeast cells, typically *S. cerevisiae*, are useful for eukaryotic genetic studies, due to the ease of targeting genetic changes by homologous recombination and the ability to easily complement genetic defects using recombinantly expressed proteins. Yeast cells are useful for identifying interacting protein components, *e.g.* through use of a two-hybrid system. In a preferred embodiment, yeast cells are useful for protein expression. Vectors of the present invention for use in yeast will typically, but not invariably, contain an origin of replication suitable for use in yeast and a selectable marker that is functional in yeast. Yeast vectors include Yeast Integrating plasmids (*e.g.*, YIp5) and Yeast Replicating plasmids (the YRp and YEp series plasmids), Yeast Centromere plasmids (the YCp series plasmids), Yeast Artificial Chromosomes (YACs) which are based on yeast linear plasmids, denoted YLp, pGPD-2, 2μ plasmids and derivatives thereof, and improved shuttle vectors such as those described in Gietz *et al.*, *Gene*, 74: 527-34 (1988) (YIplac, YEplac and YCplac). Selectable markers in yeast vectors include a variety of auxotrophic markers, the most common of which are (in *Saccharomyces cerevisiae*) URA3, HIS3, LEU2, TRP1 and LYS2, which complement specific auxotrophic mutations, such as ura3-52, his3-D1, leu2-D1, trp1-D1 and lys2-201.

Insect cells are often chosen for high efficiency protein expression. Where the host cells are from *Spodoptera frugiperda* , *e.g.*, Sf9 and Sf21 cell lines, and expresSF™ cells (Protein Sciences Corp., Meriden, CT, USA)), the vector replicative strategy is typically based upon the baculovirus life cycle. Typically, baculovirus transfer vectors are used to replace the wild-type AcMNPV polyhedrin gene with a heterologous gene of interest. Sequences that flank the polyhedrin gene in the wild-type genome are positioned 5' and 3' of the expression cassette on the transfer vectors. Following co-transfection with AcMNPV DNA, a homologous recombination event occurs between these sequences resulting in a recombinant virus carrying the gene of interest and the polyhedrin or p10 promoter. Selection can be based upon visual screening for lacZ fusion activity.

In another embodiment, the host cells may be mammalian cells, which are particularly useful for expression of proteins intended as pharmaceutical agents, and for screening of potential agonists and antagonists of a protein or a physiological pathway. Mammalian vectors intended for autonomous extrachromosomal replication will typically include a viral origin, such as the SV40 origin (for replication in cell lines expressing the large T-antigen, such as COS1 and COS7 cells), the papillomavirus origin, or the EBV origin for long term episomal replication (for use, *e.g.*, in 293-EBNA cells, which constitutively express the EBV EBNA-1 gene product and adenovirus E1A). Vectors intended for integration, and thus replication as part of the mammalian chromosome, can, but need not, include an origin of replication functional in mammalian cells, such as the SV40 origin. Vectors based upon viruses, such as adenovirus, adeno-associated virus, vaccinia virus, and various mammalian retroviruses, will typically replicate according to the viral replicative strategy. Selectable markers for use in mammalian cells include resistance to neomycin (G418), blasticidin, hygromycin and to zeocin, and selection based upon the purine salvage pathway using HAT medium.

Expression in mammalian cells can be achieved using a variety of plasmids, including pSV2, pBC12BI, and p91023, as well as lytic virus vectors (*e.g.*, vaccinia virus, adeno virus, and baculovirus), episomal virus vectors (*e.g.*, bovine papillomavirus), and retroviral vectors (*e.g.*, murine retroviruses). Useful vectors for insect cells include baculoviral vectors and pVL 941.

Plant cells can also be used for expression, with the vector replicon typically derived from a plant virus (*e.g.*, cauliflower mosaic virus, CaMV; tobacco mosaic virus, TMV) and selectable markers chosen for suitability in plants.

It is known that codon usage of different host cells may be different. For example, a plant cell and a human cell may exhibit a difference in codon preference for encoding a particular amino acid. As a result, human mRNA may not be efficiently translated in a plant, bacteria or insect host cell. Therefore, another embodiment of this invention is directed to codon optimization. The codons of the nucleic acid molecules of the invention may be modified to resemble, as much as possible, genes naturally contained within the host cell without altering the amino acid sequence encoded by the nucleic acid molecule.

Any of a wide variety of expression control sequences may be used in these vectors to express the DNA sequences of this invention. Such useful expression control sequences include the expression control sequences associated with structural genes of the foregoing expression vectors. Expression control sequences that control transcription include, *e.g.*, promoters, enhancers and transcription termination sites. Expression control sequences in eukaryotic cells that control post-transcriptional events include splice donor and acceptor sites and sequences that modify the half-life of the transcribed RNA, *e.g.*, sequences that direct poly(A) addition or binding sites for RNA-binding proteins. Expression control sequences that control translation include ribosome binding sites, sequences which direct targeted expression of the polypeptide to or within particular cellular compartments, and sequences in the 5' and 3' untranslated regions that modify the rate or efficiency of translation.

Examples of useful expression control sequences for a prokaryote, *e.g.*, *E. coli*, will include a promoter, often a phage promoter, such as phage lambda pL promoter, the trc promoter, a hybrid derived from the trp and lac promoters, the bacteriophage T7 promoter (in *E. coli* cells engineered to express the T7 polymerase), the <u>TAC</u> or <u>TRC</u> system, the major operator and promoter regions of phage lambda, the control regions of fd coat protein, or the araBAD operon. Prokaryotic expression vectors may further include transcription terminators, such as the aspA terminator, and elements that facilitate translation, such as a consensus ribosome binding site and translation termination codon, Schomer *et al.*, *Proc. Natl. Acad. Sci. USA* 83: 8506-8510 (1986).

Expression control sequences for yeast cells, typically *S. cerevisiae*, will include a yeast promoter, such as the CYC1 promoter, the GAL1 promoter, the GAL10 promoter, ADH1 promoter, the promoters of the yeast _-mating system, or the GPD promoter, and will typically have elements that facilitate transcription termination, such as the transcription termination signals from the CYC1 or ADH1 gene.

Expression vectors useful for expressing proteins in mammalian cells will include a promoter active in mammalian cells. These promoters include those derived from mammalian viruses, such as the enhancer-promoter sequences from the immediate early gene of the human cytomegalovirus (CMV), the enhancer-promoter sequences from the Rous sarcoma virus long terminal repeat (RSV LTR), the enhancer-promoter from SV40 or the early and late promoters of adenovirus. Other expression control sequences include the promoter for 3-phosphoglycerate kinase or other glycolytic enzymes, the promoters of acid phosphatase. Other expression control sequences include those from the gene comprising the CSNA of interest. Often, expression is enhanced by incorporation of polyadenylation sites, such as the late SV40 polyadenylation site and the polyadenylation signal and transcription termination sequences from the bovine growth hormone (BGH) gene, and ribosome binding sites. Furthermore, vectors can include introns, such as intron II of rabbit β-globin gene and the SV40 splice elements.

Preferred nucleic acid vectors also include a selectable or amplifiable marker gene and means for amplifying the copy number of the gene of interest. Such marker genes are well-known in the art. Nucleic acid vectors may also comprise stabilizing sequences (*e.g.*, ori- or ARS-like sequences and telomere-like sequences), or may alternatively be designed to favor directed or non-directed integration into the host cell genome. In a preferred embodiment, nucleic acid sequences of this invention are inserted in frame into an expression vector that allows high level expression of an RNA which encodes a protein comprising the encoded nucleic acid sequence of interest. Nucleic acid cloning and sequencing methods are well-known to those of skill in the art and are described in an assortment of laboratory manuals, including Sambrook (1989), *supra*, Sambrook (2000), *supra*; and Ausubel (1992), *supra*, Ausubel (1999), *supra*. Product information from manufacturers of biological, chemical and immunological reagents also provide useful information.

Expression vectors may be either constitutive or inducible. Inducible vectors include either naturally inducible promoters, such as the trc promoter, which is regulated by the lac operon, and the pL promoter, which is regulated by tryptophan, the MMTV-LTR promoter, which is inducible by dexamethasone, or can contain synthetic promoters and/or additional elements that confer inducible control on adjacent promoters. Examples of inducible synthetic promoters are the hybrid Plac/ara-1 promoter and the PLtetO-1 promoter. The PltetO-1 promoter takes advantage of the high expression levels from the PL promoter of phage lambda, but replaces the lambda repressor sites with two copies of operator 2 of the Tn10 tetracycline resistance operon, causing this promoter to be tightly repressed by the Tet repressor protein and induced in response to tetracycline (Tc) and Tc derivatives such as anhydrotetracycline. Vectors may also be inducible because they contain hormone response elements, such as the glucocorticoid response element (GRE) and the estrogen response element (ERE), which can confer hormone inducibility where vectors are used for expression in cells having the respective hormone receptors. To reduce background levels of expression, elements responsive to ecdysone, an insect hormone, can be used instead, with coexpression of the ecdysone receptor.

In one aspect of the invention, expression vectors can be designed to fuse the expressed polypeptide to small protein tags that facilitate purification and/or visualization. Tags that facilitate purification include a polyhistidine tag that facilitates purification of the fusion protein by immobilized metal affinity chromatography, for example using NiNTA resin (Qiagen Inc., Valencia, CA, USA) or TALON™ resin (cobalt immobilized affinity chromatography medium, Clontech Labs, Palo Alto, CA, USA). The fusion protein can include a chitin-binding tag and self-excising intein, permitting chitin-based purification with self-removal of the fused tag (IMPACT™ system, New England Biolabs, Inc., Beverley, MA, USA). Alternatively, the fusion protein can include a calmodulin-binding peptide tag, permitting purification by calmodulin affinity resin (Stratagene, La Jolla, CA, USA), or a specifically excisable fragment of the biotin carboxylase carrier protein, permitting purification of *in vivo* biotinylated protein using an avidin resin and subsequent tag removal (Promega, Madison, WI, USA). As another useful alternative, the proteins of the present invention can be expressed as a fusion protein with glutathione-S-transferase, the affinity and specificity of binding to glutathione permitting purification using glutathione affinity

resins, such as Glutathione-Superflow Resin (Clontech Laboratories, Palo Alto, CA, USA), with subsequent elution with free glutathione. Other tags include, for example, the Xpress epitope, detectable by anti-Xpress antibody (Invitrogen, Carlsbad, CA, USA), a myc tag, detectable by anti-myc tag antibody, the V5 epitope, detectable by anti-V5 antibody (Invitrogen, Carlsbad, CA, USA), FLAG® epitope, detectable by anti-FLAG® antibody (Stratagene, La Jolla, CA, USA), and the HA epitope.

For secretion of expressed proteins, vectors can include appropriate sequences that encode secretion signals, such as leader peptides. For example, the pSecTag2 vectors (Invitrogen, Carlsbad, CA, USA) are 5.2 kb mammalian expression vectors that carry the secretion signal from the V-J2-C region of the mouse Ig kappa-chain for efficient secretion of recombinant proteins from a variety of mammalian cell lines.

Expression vectors can also be designed to fuse proteins encoded by the heterologous nucleic acid insert to polypeptides that are larger than purification and/or identification tags. Useful fusion proteins include those that permit display of the encoded protein on the surface of a phage or cell, fusion to intrinsically fluorescent proteins, such as those that have a green fluorescent protein (GFP)-like chromophore, fusions to the IgG Fc region, and fusion proteins for use in two hybrid systems.

Vectors for phage display fuse the encoded polypeptide to, *e.g.*, the gene III protein (pIII) or gene VIII protein (pVIII) for display on the surface of filamentous phage, such as M13. *See* Barbas *et al.*, Phage Display: A Laboratory Manual, Cold Spring Harbor Laboratory Press (2001); Kay *et al.* (eds.), Phage Display of Peptides and Proteins: A Laboratory Manual, Academic Press, Inc., (1996); Abelson *et al.* (eds.), Combinatorial Chemistry (Methods in Enzymology, Vol. 267) Academic Press (1996). Vectors for yeast display, *e.g.* the pYD1 yeast display vector (Invitrogen, Carlsbad, CA, USA), use the -agglutinin yeast adhesion receptor to display recombinant protein on the surface of *S. cerevisiae*. Vectors for mammalian display, *e.g.*, the pDisplay™ vector (Invitrogen, Carlsbad, CA, USA), target recombinant proteins using an N-terminal cell surface targeting signal and a C-terminal transmembrane anchoring domain of platelet derived growth factor receptor.

A wide variety of vectors now exist that fuse proteins encoded by heterologous nucleic acids to the chromophore of the substrate-independent, intrinsically fluorescent green fluorescent protein from *Aequorea victoria* ("GFP") and its variants. The GFP-like

chromophore can be selected from GFP-like chromophores found in naturally occurring proteins, such as *A. victoria* GFP (GenBank accession number AAA27721), *Renilla reniformis* GFP, FP583 (GenBank accession no. AF168419) (DsRed), FP593 (AF272711), FP483 (AF168420), FP484 (AF168424), FP595 (AF246709), FP486 (AF168421), FP538 (AF168423), and FP506 (AF168422), and need include only so much of the native protein as is needed to retain the chromophore's intrinsic fluorescence. Methods for determining the minimal domain required for fluorescence are known in the art. *See* Li *et al., J. Biol. Chem.* 272: 28545-28549 (1997). Alternatively, the GFP-like chromophore can be selected from GFP-like chromophores modified from those found in nature. The methods for engineering such modified GFP-like chromophores and testing them for fluorescence activity, both alone and as part of protein fusions, are well-known in the art. *See* Heim *et al., Curr. Biol.* 6: 178-182 (1996) and Palm *et al., Methods Enzymol.* 302: 378-394 (1999), incorporated herein by reference in its entirety. A variety of such modified chromophores are now commercially available and can readily be used in the fusion proteins of the present invention. These include EGFP ("enhanced GFP"), EBFP ("enhanced blue fluorescent protein"), BFP2, EYFP ("enhanced yellow fluorescent protein"), ECFP ("enhanced cyan fluorescent protein") or Citrine. EGFP (*see, e.g,* Cormack *et al., Gene* 173: 33–38 (1996); United States Patent Nos. 6,090,919 and 5,804,387) is found on a variety of vectors, both plasmid and viral, which are available commercially (Clontech Labs, Palo Alto, CA, USA); EBFP is optimized for expression in mammalian cells whereas BFP2, which retains the original jellyfish codons, can be expressed in bacteria (*see, e.g,.* Heim *et al., Curr. Biol.* 6: 178-182 (1996) and Cormack *et al., Gene* 173: 33-38 (1996)). Vectors containing these blue-shifted variants are available from Clontech Labs (Palo Alto, CA, USA). Vectors containing EYFP, ECFP (*see, e.g.,* Heim *et al., Curr. Biol.* 6: 178-182 (1996); Miyawaki *et al., Nature* 388: 882-887 (1997)) and Citrine (*see, e.g.,* Heikal *et al., Proc. Natl. Acad. Sci. USA* 97: 11996-12001 (2000)) are also available from Clontech Labs. The GFP-like chromophore can also be drawn from other modified GFPs, including those described in U.S. Patents 6,124,128; 6,096,865; 6,090,919; 6,066,476; 6,054,321; 6,027,881; 5,968,750; 5,874,304; 5,804,387; 5,777,079; 5,741,668; and 5,625,048, the disclosures of which are incorporated herein by reference in their entireties. *See* also Conn (ed.), Green Fluorescent Protein (Methods in

Enzymology, Vol. 302), Academic Press, Inc. (1999). The GFP-like chromophore of each of these GFP variants can usefully be included in the fusion proteins of the present invention.

Fusions to the IgG Fc region increase serum half life of protein pharmaceutical products through interaction with the FcRn receptor (also denominated the FcRp receptor and the Brambell receptor, FcRb), further described in International Patent Application Nos. WO 97/43316, WO 97/34631, WO 96/32478, WO 96/18412.

For long-term, high-yield recombinant production of the proteins, protein fusions, and protein fragments of the present invention, stable expression is preferred. Stable expression is readily achieved by integration into the host cell genome of vectors having selectable markers, followed by selection of these integrants. Vectors such as pUB6/V5-His A, B, and C (Invitrogen, Carlsbad, CA, USA) are designed for high-level stable expression of heterologous proteins in a wide range of mammalian tissue types and cell lines. pUB6/V5-His uses the promoter/enhancer sequence from the human ubiquitin C gene to drive expression of recombinant proteins: expression levels in 293, CHO, and NIH3T3 cells are comparable to levels from the CMV and human EF-1a promoters. The bsd gene permits rapid selection of stably transfected mammalian cells with the potent antibiotic blasticidin.

Replication incompetent retroviral vectors, typically derived from Moloney murine leukemia virus, also are useful for creating stable transfectants having integrated provirus. The highly efficient transduction machinery of retroviruses, coupled with the availability of a variety of packaging cell lines such as RetroPack™ PT 67, EcoPack2™-293, AmphoPack-293, and GP2-293 cell lines (all available from Clontech Laboratories, Palo Alto, CA, USA), allow a wide host range to be infected with high efficiency; varying the multiplicity of infection readily adjusts the copy number of the integrated provirus.

Of course, not all vectors and expression control sequences will function equally well to express the nucleic acid sequences of this invention. Neither will all hosts function equally well with the same expression system. However, one of skill in the art may make a selection among these vectors, expression control sequences and hosts without undue experimentation and without departing from the scope of this invention. For example, in selecting a vector, the host must be considered because the vector must

be replicated in it. The vector's copy number, the ability to control that copy number, the ability to control integration, if any, and the expression of any other proteins encoded by the vector, such as antibiotic or other selection markers, should also be considered. The present invention further includes host cells comprising the vectors of the present invention, either present episomally within the cell or integrated, in whole or in part, into the host cell chromosome. Among other considerations, some of which are described above, a host cell strain may be chosen for its ability to process the expressed protein in the desired fashion. Such post-translational modifications of the polypeptide include, but are not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation, and acylation, and it is an aspect of the present invention to provide CSPs with such post-translational modifications.

Polypeptides of the invention may be post-translationally modified. Post-translational modifications include phosphorylation of amino acid residues serine, threonine and/or tyrosine, N-linked and/or O-linked glycosylation, methylation, acetylation, prenylation, methylation, acetylation, arginylation, ubiquination and racemization. One may determine whether a polypeptide of the invention is likely to be post-translationally modified by analyzing the sequence of the polypeptide to determine if there are peptide motifs indicative of sites for post-translational modification. There are a number of computer programs that permit prediction of post-translational modifications. See, e.g., www.expasy.org (accessed August 31, 2001), which includes PSORT, for prediction of protein sorting signals and localization sites, SignalP, for prediction of signal peptide cleavage sites, MITOPROT and Predotar, for prediction of mitochondrial targeting sequences, NetOGlyc, for prediction of type O-glycosylation sites in mammalian proteins, big-PI Predictor and DGPI, for prediction of prenylation-anchor and cleavage sites, and NetPhos, for prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins. Other computer programs, such as those included in GCG, also may be used to determine post-translational modification peptide motifs.

General examples of types of post-translational modifications may be found in web sites such as the Delta Mass database http://www.abrf.org/ABRF/Research Committees/deltamass/deltamass.html (accessed October 19, 2001); "GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological

sources" Cooper et al. Nucleic Acids Res. 29; 332-335 (2001) and http://www.glycosuite.com/ (accessed October 19, 2001); "O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins" Gupta et al. Nucleic Acids Research, 27: 370-372 (1999) and http://www.cbs.dtu.dk/databases/OGLYCBASE/ (accessed October 19, 2001); "PhosphoBase, a database of phosphorylation sites: release 2.0.", Kreegipuu et al. Nucleic Acids Res 27(1):237-239 (1999) and http://www.cbs.dtu.dk/ databases/PhosphoBase/ (accessed October 19, 2001); or http://pir.georgetown.edu/ pirwww/search/textresid.html (accessed October 19, 2001).

Tumorigenesis is often accompanied by alterations in the post-translational modifications of proteins. Thus, in another embodiment, the invention provides polypeptides from cancerous cells or tissues that have altered post-translational modifications compared to the post-translational modifications of polypeptides from normal cells or tissues. A number of altered post-translational modifications are known. One common alteration is a change in phosphorylation state, wherein the polypeptide from the cancerous cell or tissue is hyperphosphorylated or hypophosphorylated compared to the polypeptide from a normal tissue, or wherein the polypeptide is phosphorylated on different residues than the polypeptide from a normal cell. Another common alteration is a change in glycosylation state, wherein the polypeptide from the cancerous cell or tissue has more or less glycosylation than the polypeptide from a normal tissue, and/or wherein the polypeptide from the cancerous cell or tissue has a different type of glycosylation than the polypeptide from a noncancerous cell or tissue. Changes in glycosylation may be critical because carbohydrate-protein and carbohydrate-carbohydrate interactions are important in cancer cell progression, dissemination and invasion. See, e.g., Barchi, *Curr. Pharm. Des.* 6: 485-501 (2000), Verma, *Cancer Biochem. Biophys.* 14: 151-162 (1994) and Dennis et al., *Bioessays* 5: 412-421 (1999).

Another post-translational modification that may be altered in cancer cells is prenylation. Prenylation is the covalent attachment of a hydrophobic prenyl group (either farnesyl or geranylgeranyl) to a polypeptide. Prenylation is required for localizing a protein to a cell membrane and is often required for polypeptide function. For instance, the Ras superfamily of GTPase signaling proteins must be prenylated for function in a cell. See, e.g., Prendergast et al., *Semin. Cancer Biol.* 10: 443-452 (2000) and Khwaja et al., *Lancet* 355: 741-744 (2000).

Other post-translation modifications that may be altered in cancer cells include, without limitation, polypeptide methylation, acetylation, arginylation or racemization of amino acid residues. In these cases, the polypeptide from the cancerous cell may exhibit either increased or decreased amounts of the post-translational modification compared to the corresponding polypeptides from noncancerous cells.

Other polypeptide alterations in cancer cells include abnormal polypeptide cleavage of proteins and aberrant protein-protein interactions. Abnormal polypeptide cleavage may be cleavage of a polypeptide in a cancerous cell that does not usually occur in a normal cell, or a lack of cleavage in a cancerous cell, wherein the polypeptide is cleaved in a normal cell. Aberrant protein-protein interactions may be either covalent cross-linking or non-covalent binding between proteins that do not normally bind to each other. Alternatively, in a cancerous cell, a protein may fail to bind to another protein to which it is bound in a noncancerous cell. Alterations in cleavage or in protein-protein interactions may be due to over- or underproduction of a polypeptide in a cancerous cell compared to that in a normal cell, or may be due to alterations in post-translational modifications (see above) of one or more proteins in the cancerous cell. See, e.g., Henschen-Edman, *Ann. N.Y. Acad. Sci.* 936: 580-593 (2001).

Alterations in polypeptide post-translational modifications, as well as changes in polypeptide cleavage and protein-protein interactions, may be determined by any method known in the art. For instance, alterations in phosphorylation may be determined by using anti-phosphoserine, anti-phosphothreonine or anti-phosphotyrosine antibodies or by amino acid analysis. Glycosylation alterations may be determined using antibodies specific for different sugar residues, by carbohydrate sequencing, or by alterations in the size of the glycoprotein, which can be determined by, e.g., SDS polyacrylamide gel electrophoresis (PAGE). Other alterations of post-translational modifications, such as prenylation, racemization, methylation, acetylation and arginylation, may be determined by chemical analysis, protein sequencing, amino acid analysis, or by using antibodies specific for the particular post-translational modifications. Changes in protein-protein interactions and in polypeptide cleavage may be analyzed by any method known in the art including, without limitation, non-denaturing PAGE (for non-covalent protein-protein interactions), SDS PAGE (for covalent protein-protein interactions and protein cleavage), chemical cleavage, protein sequencing or immunoassays.

In another embodiment, the invention provides polypeptides that have been post-translationally modified. In one embodiment, polypeptides may be modified enzymatically or chemically, by addition or removal of a post-translational modification. For example, a polypeptide may be glycosylated or deglycosylated enzymatically. Similarly, polypeptides may be phosphorylated using a purified kinase, such as a MAP kinase (e.g, p38, ERK, or JNK) or a tyrosine kinase (e.g., Src or erbB2). A polypeptide may also be modified through synthetic chemistry. Alternatively, one may isolate the polypeptide of interest from a cell or tissue that expresses the polypeptide with the desired post-translational modification. In another embodiment, a nucleic acid molecule encoding the polypeptide of interest is introduced into a host cell that is capable of post-translationally modifying the encoded polypeptide in the desired fashion. If the polypeptide does not contain a motif for a desired post-translational modification, one may alter the post-translational modification by mutating the nucleic acid sequence of a nucleic acid molecule encoding the polypeptide so that it contains a site for the desired post-translational modification. Amino acid sequences that may be post-translationally modified are known in the art. See, e.g., the programs described above on the website www.expasy.org. The nucleic acid molecule is then be introduced into a host cell that is capable of post-translationally modifying the encoded polypeptide. Similarly, one may delete sites that are post-translationally modified by either mutating the nucleic acid sequence so that the encoded polypeptide does not contain the post-translational modification motif, or by introducing the native nucleic acid molecule into a host cell that is not capable of post-translationally modifying the encoded polypeptide.

In selecting an expression control sequence, a variety of factors should also be considered. These include, for example, the relative strength of the sequence, its controllability, and its compatibility with the nucleic acid sequence of this invention, particularly with regard to potential secondary structures. Unicellular hosts should be selected by consideration of their compatibility with the chosen vector, the toxicity of the product coded for by the nucleic acid sequences of this invention, their secretion characteristics, their ability to fold the polypeptide correctly, their fermentation or culture requirements, and the ease of purification from them of the products coded for by the nucleic acid sequences of this invention.

The recombinant nucleic acid molecules and more particularly, the expression vectors of this invention may be used to express the polypeptides of this invention as recombinant polypeptides in a heterologous host cell. The polypeptides of this invention may be full-length or less than full-length polypeptide fragments recombinantly expressed from the nucleic acid sequences according to this invention. Such polypeptides include analogs, derivatives and muteins that may or may not have biological activity.

Vectors of the present invention will also often include elements that permit *in vitro* transcription of RNA from the inserted heterologous nucleic acid. Such vectors typically include a phage promoter, such as that from T7, T3, or SP6, flanking the nucleic acid insert. Often two different such promoters flank the inserted nucleic acid, permitting separate *in vitro* production of both sense and antisense strands.

Transformation and other methods of introducing nucleic acids into a host cell (*e.g.*, conjugation, protoplast transformation or fusion, transfection, electroporation, liposome delivery, membrane fusion techniques, high velocity DNA-coated pellets, viral infection and protoplast fusion) can be accomplished by a variety of methods which are well-known in the art (*See*, for instance, Ausubel, *supra*, and Sambrook *et al.*, *supra*). Bacterial, yeast, plant or mammalian cells are transformed or transfected with an expression vector, such as a plasmid, a cosmid, or the like, wherein the expression vector comprises the nucleic acid of interest. Alternatively, the cells may be infected by a viral expression vector comprising the nucleic acid of interest. Depending upon the host cell, vector, and method of transformation used, transient or stable expression of the polypeptide will be constitutive or inducible. One having ordinary skill in the art will be able to decide whether to express a polypeptide transiently or stably, and whether to express the protein constitutively or inducibly.

A wide variety of unicellular host cells are useful in expressing the DNA sequences of this invention. These hosts may include well-known eukaryotic and prokaryotic hosts, such as strains of, fungi, yeast, insect cells such as *Spodoptera frugiperda* (SF9), animal cells such as CHO, as well as plant cells in tissue culture. Representative examples of appropriate host cells include, but are not limited to, bacterial cells, such as *E. coli*, *Caulobacter crescentus*, *Streptomyces* species, and *Salmonella typhimurium*; yeast cells, such as *Saccharomyces cerevisiae, Schizosaccharomyces*

pombe, *Pichia pastoris*, *Pichia methanolica*; insect cell lines, such as those from *Spodoptera frugiperda*, *e.g.*, Sf9 and Sf21 cell lines, and expresSF™ cells (Protein Sciences Corp., Meriden, CT, USA), *Drosophila* S2 cells, and *Trichoplusia ni* High Five® Cells (Invitrogen, Carlsbad, CA, USA); and mammalian cells. Typical mammalian cells include BHK cells, BSC 1 cells, BSC 40 cells, BMT 10 cells, VERO cells, COS1 cells, COS7 cells, Chinese hamster ovary (CHO) cells, 3T3 cells, NIH 3T3 cells, 293 cells, HEPG2 cells, HeLa cells, L cells, MDCK cells, HEK293 cells, WI38 cells, murine ES cell lines (*e.g.*, from strains 129/SV, C57/BL6, DBA-1, 129/SVJ), K562 cells, Jurkat cells, and BW5147 cells. Other mammalian cell lines are well-known and readily available from the American Type Culture Collection (ATCC) (Manassas, VA, USA) and the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository at the Coriell Cell Repositories (Camden, NJ, USA). Cells or cell lines derived from colon are particularly preferred because they may provide a more native post-translational processing. Particularly preferred are human colon cells.

Particular details of the transfection, expression and purification of recombinant proteins are well documented and are understood by those of skill in the art. Further details on the various technical aspects of each of the steps used in recombinant production of foreign genes in bacterial cell expression systems can be found in a number of texts and laboratory manuals in the art. *See, e.g.*, Ausubel (1992), *supra*, Ausubel (1999), *supra*, Sambrook (1989), *supra*, and Sambrook (2001), *supra*, herein incorporated by reference.

Methods for introducing the vectors and nucleic acids of the present invention into the host cells are well-known in the art; the choice of technique will depend primarily upon the specific vector to be introduced and the host cell chosen.

Nucleic acid molecules and vectors may be introduced into prokaryotes, such as *E. coli*, in a number of ways. For instance, phage lambda vectors will typically be packaged using a packaging extract (*e.g.*, Gigapack® packaging extract, Stratagene, La Jolla, CA, USA), and the packaged virus used to infect *E. coli*.

Plasmid vectors will typically be introduced into chemically competent or electrocompetent bacterial cells. *E. coli* cells can be rendered chemically competent by treatment, *e.g.*, with $CaCl_2$, or a solution of Mg^{2+}, Mn^{2+}, Ca^{2+}, Rb^{+} or K^{+}, dimethyl sulfoxide, dithiothreitol, and hexamine cobalt (III), Hanahan, *J. Mol. Biol.* 166(4):557-80

(1983), and vectors introduced by heat shock. A wide variety of chemically competent strains are also available commercially (*e.g.*, Epicurian Coli® XL10-Gold® Ultracompetent Cells (Stratagene, La Jolla, CA, USA); DH5 competent cells (Clontech Laboratories, Palo Alto, CA, USA); and TOP10 Chemically Competent E. coli Kit (Invitrogen, Carlsbad, CA, USA)). Bacterial cells can be rendered electrocompetent, that is, competent to take up exogenous DNA by electroporation, by various pre-pulse treatments; vectors are introduced by electroporation followed by subsequent outgrowth in selected media. An extensive series of protocols is provided online in <u>Electroprotocols</u> (BioRad, Richmond, CA, USA) (http://www.biorad.com/LifeScience/pdf/ New_Gene_Pulser.pdf).

Vectors can be introduced into yeast cells by spheroplasting, treatment with lithium salts, electroporation, or protoplast fusion. Spheroplasts are prepared by the action of hydrolytic enzymes such as snail-gut extract, usually denoted Glusulase, or Zymolyase, an enzyme from *Arthrobacter luteus*, to remove portions of the cell wall in the presence of osmotic stabilizers, typically 1 M sorbitol. DNA is added to the spheroplasts, and the mixture is co-precipitated with a solution of polyethylene glycol (PEG) and Ca^{2+}. Subsequently, the cells are resuspended in a solution of sorbitol, mixed with molten agar and then layered on the surface of a selective plate containing sorbitol.

For lithium-mediated transformation, yeast cells are treated with lithium acetate, which apparently permeabilizes the cell wall, DNA is added and the cells are co-precipitated with PEG. The cells are exposed to a brief heat shock, washed free of PEG and lithium acetate, and subsequently spread on plates containing ordinary selective medium. Increased frequencies of transformation are obtained by using specially-prepared single-stranded carrier DNA and certain organic solvents. Schiestl *et al.*, *Curr. Genet.* 16(5-6): 339-46 (1989).

For electroporation, freshly-grown yeast cultures are typically washed, suspended in an osmotic protectant, such as sorbitol, mixed with DNA, and the cell suspension pulsed in an electroporation device. Subsequently, the cells are spread on the surface of plates containing selective media. Becker *et al.*, *Methods Enzymol.* 194: 182-187 (1991). The efficiency of transformation by electroporation can be increased over 100-fold by using PEG, single-stranded carrier DNA and cells that are in late log-phase of growth. Larger constructs, such as YACs, can be introduced by protoplast fusion.

Mammalian and insect cells can be directly infected by packaged viral vectors, or transfected by chemical or electrical means. For chemical transfection, DNA can be coprecipitated with $CaPO_4$ or introduced using liposomal and nonliposomal lipid-based agents. Commercial kits are available for $CaPO_4$ transfection (CalPhos™ Mammalian Transfection Kit, Clontech Laboratories, Palo Alto, CA, USA), and lipid-mediated transfection can be practiced using commercial reagents, such as LIPOFECTAMINE™ 2000, LIPOFECTAMINE™ Reagent, CELLFECTIN® Reagent, and LIPOFECTIN® Reagent (Invitrogen, Carlsbad, CA, USA), DOTAP Liposomal Transfection Reagent, FuGENE 6, X-tremeGENE Q2, DOSPER, (Roche Molecular Biochemicals, Indianapolis, IN USA), Effectene™, PolyFect®, Superfect® (Qiagen, Inc., Valencia, CA, USA). Protocols for electroporating mammalian cells can be found online in Electroprotocols (Bio-Rad, Richmond, CA, USA) (http://www.bio-rad.com/LifeScience/pdf/ New_Gene_Pulser.pdf); Norton *et al.* (eds.), <u>Gene Transfer Methods: Introducing DNA into Living Cells and Organisms</u>, BioTechniques Books, Eaton Publishing Co. (2000); incorporated herein by reference in its entirety. Other transfection techniques include transfection by particle bombardment and microinjection. *See, e.g.*, Cheng *et al., Proc. Natl. Acad. Sci. USA* 90(10): 4455-9 (1993); Yang *et al., Proc. Natl. Acad. Sci. USA* 87(24): 9568-72 (1990).

Production of the recombinantly produced proteins of the present invention can optionally be followed by purification.

Purification of recombinantly expressed proteins is now well by those skilled in the art. *See, e.g.*, Thorner *et al.* (eds.), <u>Applications of Chimeric Genes and Hybrid Proteins, Part A: Gene Expression and Protein Purification</u> (Methods in Enzymology, Vol. 326), Academic Press (2000); Harbin (ed.), <u>Cloning, Gene Expression and Protein Purification : Experimental Procedures and Process Rationale</u>, Oxford Univ. Press (2001); Marshak *et al.*, <u>Strategies for Protein Purification and Characterization: A Laboratory Course Manual</u>, Cold Spring Harbor Laboratory Press (1996); and Roe (ed.), <u>Protein Purification Applications</u>, Oxford University Press (2001); the disclosures of which are incorporated herein by reference in their entireties, and thus need not be detailed here.

Briefly, however, if purification tags have been fused through use of an expression vector that appends such tags, purification can be effected, at least in part, by

means appropriate to the tag, such as use of immobilized metal affinity chromatography for polyhistidine tags. Other techniques common in the art include ammonium sulfate fractionation, immunoprecipitation, fast protein liquid chromatography (FPLC), high performance liquid chromatography (HPLC), and preparative gel electrophoresis.

Polypeptides

Another object of the invention is to provide polypeptides encoded by the nucleic acid molecules of the instant invention. In a preferred embodiment, the polypeptide is a colon specific polypeptide (CSP). In an even more preferred embodiment, the polypeptide is derived from a polypeptide comprising the amino acid sequence of SEQ ID NO: 101 through 176. A polypeptide as defined herein may be produced recombinantly, as discussed *supra*, may be isolated from a cell that naturally expresses the protein, or may be chemically synthesized following the teachings of the specification and using methods well-known to those having ordinary skill in the art.

In another aspect, the polypeptide may comprise a fragment of a polypeptide, wherein the fragment is as defined herein. In a preferred embodiment, the polypeptide fragment is a fragment of a CSP. In a more preferred embodiment, the fragment is derived from a polypeptide comprising the amino acid sequence of SEQ ID NO: 101 through 176. A polypeptide that comprises only a fragment of an entire CSP may or may not be a polypeptide that is also a CSP. For instance, a full-length polypeptide may be colon-specific, while a fragment thereof may be found in other tissues as well as in colon. A polypeptide that is not a CSP, whether it is a fragment, analog, mutein, homologous protein or derivative, is nevertheless useful, especially for immunizing animals to prepare anti-CSP antibodies. However, in a preferred embodiment, the part or fragment is a CSP. Methods of determining whether a polypeptide is a CSP are described *infra*.

Fragments of at least 6 contiguous amino acids are useful in mapping B cell and T cell epitopes of the reference protein. *See, e.g.*, Geysen *et al.*, *Proc. Natl. Acad. Sci. USA* 81: 3998-4002 (1984) and U.S. Patents 4,708,871 and 5,595,915, the disclosures of which are incorporated herein by reference in their entireties. Because the fragment need not itself be immunogenic, part of an immunodominant epitope, nor even recognized by native antibody, to be useful in such epitope mapping, all fragments of at least 6 amino acids of the proteins of the present invention have utility in such a study.

Fragments of at least 8 contiguous amino acids, often at least 15 contiguous amino acids, are useful as immunogens for raising antibodies that recognize the proteins of the present invention. *See, e.g.*, Lerner, Nature 299: 592-596 (1982); Shinnick *et al.*, *Annu. Rev. Microbiol.* 37: 425-46 (1983); Sutcliffe *et al.*, *Science* 219: 660-6 (1983), the disclosures of which are incorporated herein by reference in their entireties. As further described in the above-cited references, virtually all 8-mers, conjugated to a carrier, such as a protein, prove immunogenic, meaning that they are capable of eliciting antibody for the conjugated peptide; accordingly, all fragments of at least 8 amino acids of the proteins of the present invention have utility as immunogens.

Fragments of at least 8, 9, 10 or 12 contiguous amino acids are also useful as competitive inhibitors of binding of the entire protein, or a portion thereof, to antibodies (as in epitope mapping), and to natural binding partners, such as subunits in a multimeric complex or to receptors or ligands of the subject protein; this competitive inhibition permits identification and separation of molecules that bind specifically to the protein of interest, U.S. Patents 5,539,084 and 5,783,674, incorporated herein by reference in their entireties.

The protein, or protein fragment, of the present invention is thus at least 6 amino acids in length, typically at least 8, 9, 10 or 12 amino acids in length, and often at least 15 amino acids in length. Often, the protein of the present invention, or fragment thereof, is at least 20 amino acids in length, even 25 amino acids, 30 amino acids, 35 amino acids, or 50 amino acids or more in length. Of course, larger fragments having at least 75 amino acids, 100 amino acids, or even 150 amino acids are also useful, and at times preferred.

One having ordinary skill in the art can produce fragments of a polypeptide by truncating the nucleic acid molecule, *e.g.*, a CSNA, encoding the polypeptide and then expressing it recombinantly. Alternatively, one can produce a fragment by chemically synthesizing a portion of the full-length polypeptide. One may also produce a fragment by enzymatically cleaving either a recombinant polypeptide or an isolated naturally-occurring polypeptide. Methods of producing polypeptide fragments are well-known in the art. *See, e.g.*, Sambrook (1989), *supra;* Sambrook (2001), *supra*; Ausubel (1992), *supra*; and Ausubel (1999), *supra*. In one embodiment, a polypeptide comprising only a fragment of polypeptide of the invention, preferably a CSP, may be produced by

chemical or enzymatic cleavage of a polypeptide. In a preferred embodiment, a polypeptide fragment is produced by expressing a nucleic acid molecule encoding a fragment of the polypeptide, preferably a CSP, in a host cell.

By "polypeptides" as used herein it is also meant to be inclusive of mutants, fusion proteins, homologous proteins and allelic variants of the polypeptides specifically exemplified.

A mutant protein, or mutein, may have the same or different properties compared to a naturally-occurring polypeptide and comprises at least one amino acid insertion, duplication, deletion, rearrangement or substitution compared to the amino acid sequence of a native protein. Small deletions and insertions can often be found that do not alter the function of the protein. In one embodiment, the mutein may or may not be colon-specific. In a preferred embodiment, the mutein is colon-specific. In a preferred embodiment, the mutein is a polypeptide that comprises at least one amino acid insertion, duplication, deletion, rearrangement or substitution compared to the amino acid sequence of SEQ ID NO: 101 through 176. In a more preferred embodiment, the mutein is one that exhibits at least 50% sequence identity, more preferably at least 60% sequence identity, even more preferably at least 70%, yet more preferably at least 80% sequence identity to a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176. In yet a more preferred embodiment, the mutein exhibits at least 85%, more preferably 90%, even more preferably 95% or 96%, and yet more preferably at least 97%, 98%, 99% or 99.5% sequence identity to a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176.

A mutein may be produced by isolation from a naturally-occurring mutant cell, tissue or organism. A mutein may be produced by isolation from a cell, tissue or organism that has been experimentally mutagenized. Alternatively, a mutein may be produced by chemical manipulation of a polypeptide, such as by altering the amino acid residue to another amino acid residue using synthetic or semi-synthetic chemical techniques. In a preferred embodiment, a mutein may be produced from a host cell comprising an altered nucleic acid molecule compared to the naturally-occurring nucleic acid molecule. For instance, one may produce a mutein of a polypeptide by introducing one or more mutations into a nucleic acid sequence of the invention and then expressing it recombinantly. These mutations may be targeted, in which particular encoded amino

acids are altered, or may be untargeted, in which random encoded amino acids within the polypeptide are altered. Muteins with random amino acid alterations can be screened for a particular biological activity or property, particularly whether the polypeptide is colon-specific, as described below. Multiple random mutations can be introduced into the gene by methods well-known to the art, *e.g.*, by error-prone PCR, shuffling, oligonucleotide-directed mutagenesis, assembly PCR, sexual PCR mutagenesis, *in vivo* mutagenesis, cassette mutagenesis, recursive ensemble mutagenesis, exponential ensemble mutagenesis and site-specific mutagenesis. Methods of producing muteins with targeted or random amino acid alterations are well-known in the art. *See, e.g.*, Sambrook (1989), *supra;* Sambrook (2001), *supra*; Ausubel (1992), *supra*; and Ausubel (1999), U.S. Patent 5,223,408, and the references discussed *supra*, each herein incorporated by reference.

By "polypeptide" as used herein it is also meant to be inclusive of polypeptides homologous to those polypeptides exemplified herein. In a preferred embodiment, the polypeptide is homologous to a CSP. In an even more preferred embodiment, the polypeptide is homologous to a CSP selected from the group having an amino acid sequence of SEQ ID NO: 101 through 176. In a preferred embodiment, the homologous polypeptide is one that exhibits significant sequence identity to a CSP. In a more preferred embodiment, the polypeptide is one that exhibits significant sequence identity to an comprising an amino acid sequence of SEQ ID NO: 101 through 176. In an even more preferred embodiment, the homologous polypeptide is one that exhibits at least 50% sequence identity, more preferably at least 60% sequence identity, even more preferably at least 70%, yet more preferably at least 80% sequence identity to a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176. In a yet more preferred embodiment, the homologous polypeptide is one that exhibits at least 85%, more preferably 90%, even more preferably 95% or 96%, and yet more preferably at least 97% or 98% sequence identity to a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176. In another preferred embodiment, the homologous polypeptide is one that exhibits at least 99%, more preferably 99.5%, even more preferably 99.6%, 99.7%, 99.8% or 99.9% sequence identity to a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176. In a preferred embodiment, the amino acid substitutions are conservative amino acid substitutions as discussed above.

In another embodiment, the homologous polypeptide is one that is encoded by a nucleic acid molecule that selectively hybridizes to a CSNA. In a preferred embodiment, the homologous polypeptide is encoded by a nucleic acid molecule that hybridizes to a CSNA under low stringency, moderate stringency or high stringency conditions, as defined herein. In a more preferred embodiment, the CSNA is selected from the group consisting of SEQ ID NO: 1 through 100. In another preferred embodiment, the homologous polypeptide is encoded by a nucleic acid molecule that hybridizes to a nucleic acid molecule that encodes a CSP under low stringency, moderate stringency or high stringency conditions, as defined herein. In a more preferred embodiment, the CSP is selected from the group consisting of SEQ ID NO: 101 through 176.

The homologous polypeptide may be a naturally-occurring one that is derived from another species, especially one derived from another primate, such as chimpanzee, gorilla, rhesus macaque, baboon or gorilla, wherein the homologous polypeptide comprises an amino acid sequence that exhibits significant sequence identity to that of SEQ ID NO: 101 through 176. The homologous polypeptide may also be a naturally-occurring polypeptide from a human, when the CSP is a member of a family of polypeptides. The homologous polypeptide may also be a naturally-occurring polypeptide derived from a non-primate, mammalian species, including without limitation, domesticated species, *e.g.*, dog, cat, mouse, rat, rabbit, guinea pig, hamster, cow, horse, goat or pig. The homologous polypeptide may also be a naturally-occurring polypeptide derived from a non-mammalian species, such as birds or reptiles. The naturally-occurring homologous protein may be isolated directly from humans or other species. Alternatively, the nucleic acid molecule encoding the naturally-occurring homologous polypeptide may be isolated and used to express the homologous polypeptide recombinantly. In another embodiment, the homologous polypeptide may be one that is experimentally produced by random mutation of a nucleic acid molecule and subsequent expression of the nucleic acid molecule. In another embodiment, the homologous polypeptide may be one that is experimentally produced by directed mutation of one or more codons to alter the encoded amino acid of a CSP. Further, the homologous protein may or may not encode polypeptide that is a CSP. However, in a preferred embodiment, the homologous polypeptide encodes a polypeptide that is a CSP.

Relatedness of proteins can also be characterized using a second functional test, the ability of a first protein competitively to inhibit the binding of a second protein to an antibody. It is, therefore, another aspect of the present invention to provide isolated proteins not only identical in sequence to those described with particularity herein, but also to provide isolated proteins ("cross-reactive proteins") that competitively inhibit the binding of antibodies to all or to a portion of various of the isolated polypeptides of the present invention. Such competitive inhibition can readily be determined using immunoassays well-known in the art.

As discussed above, single nucleotide polymorphisms (SNPs) occur frequently in eukaryotic genomes, and the sequence determined from one individual of a species may differ from other allelic forms present within the population. Thus, by "polypeptide" as used herein it is also meant to be inclusive of polypeptides encoded by an allelic variant of a nucleic acid molecule encoding a CSP. In a preferred embodiment, the polypeptide is encoded by an allelic variant of a gene that encodes a polypeptide having the amino acid sequence selected from the group consisting of SEQ ID NO: 101 through 176. In a yet more preferred embodiment, the polypeptide is encoded by an allelic variant of a gene that has the nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through 100.

In another embodiment, the invention provides polypeptides which comprise derivatives of a polypeptide encoded by a nucleic acid molecule according to the instant invention. In a preferred embodiment, the polypeptide is a CSP. In a preferred embodiment, the polypeptide has an amino acid sequence selected from the group consisting of SEQ ID NO: 101 through 176, or is a mutein, allelic variant, homologous protein or fragment thereof. In a preferred embodiment, the derivative has been acetylated, carboxylated, phosphorylated, glycosylated or ubiquitinated. In another preferred embodiment, the derivative has been labeled with, *e.g.*, radioactive isotopes such as ^{125}I, ^{32}P, ^{35}S, and ^{3}H. In another preferred embodiment, the derivative has been labeled with fluorophores, chemiluminescent agents, enzymes, and antiligands that can serve as specific binding pair members for a labeled ligand.

Polypeptide modifications are well-known to those of skill and have been described in great detail in the scientific literature. Several particularly common modifications, glycosylation, lipid attachment, sulfation, gamma-carboxylation of

glutamic acid residues, hydroxylation and ADP-ribosylation, for instance, are described in most basic texts, such as, for instance Creighton, Protein Structure and Molecular Properties, 2nd ed., W. H. Freeman and Company (1993). Many detailed reviews are available on this subject, such as, for example, those provided by Wold, in Johnson (ed.), Posttranslational Covalent Modification of Proteins, pgs. 1-12, Academic Press (1983); Seifter *et al.*, *Meth. Enzymol.* 182: 626-646 (1990) and Rattan *et al.*, *Ann. N.Y. Acad. Sci.* 663: 48-62 (1992).

It will be appreciated, as is well-known and as noted above, that polypeptides are not always entirely linear. For instance, polypeptides may be branched as a result of ubiquitination, and they may be circular, with or without branching, generally as a result of posttranslation events, including natural processing event and events brought about by human manipulation which do not occur naturally. Circular, branched and branched circular polypeptides may be synthesized by non-translation natural process and by entirely synthetic methods, as well. Modifications can occur anywhere in a polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. In fact, blockage of the amino or carboxyl group in a polypeptide, or both, by a covalent modification, is common in naturally occurring and synthetic polypeptides and such modifications may be present in polypeptides of the present invention, as well. For instance, the amino terminal residue of polypeptides made in *E. coli*, prior to proteolytic processing, almost invariably will be N-formylmethionine.

Useful post-synthetic (and post-translational) modifications include conjugation to detectable labels, such as fluorophores. A wide variety of amine-reactive and thiol-reactive fluorophore derivatives have been synthesized that react under nondenaturing conditions with N-terminal amino groups and epsilon amino groups of lysine residues, on the one hand, and with free thiol groups of cysteine residues, on the other.

Kits are available commercially that permit conjugation of proteins to a variety of amine-reactive or thiol-reactive fluorophores: Molecular Probes, Inc. (Eugene, OR, USA), *e.g.*, offers kits for conjugating proteins to Alexa Fluor 350, Alexa Fluor 430, Fluorescein-EX, Alexa Fluor 488, Oregon Green 488, Alexa Fluor 532, Alexa Fluor 546, Alexa Fluor 546, Alexa Fluor 568, Alexa Fluor 594, and Texas Red-X.

A wide variety of other amine-reactive and thiol-reactive fluorophores are available commercially (Molecular Probes, Inc., Eugene, OR, USA), including Alexa

Fluor® 350, Alexa Fluor® 488, Alexa Fluor® 532, Alexa Fluor® 546, Alexa Fluor® 568, Alexa Fluor® 594, Alexa Fluor® 647 (monoclonal antibody labeling kits available from Molecular Probes, Inc., Eugene, OR, USA), BODIPY dyes, such as BODIPY 493/503, BODIPY FL, BODIPY R6G, BODIPY 530/550, BODIPY TMR, BODIPY 558/568, BODIPY 558/568, BODIPY 564/570, BODIPY 576/589, BODIPY 581/591, BODIPY TR, BODIPY 630/650, BODIPY 650/665, Cascade Blue, Cascade Yellow, Dansyl, lissamine rhodamine B, Marina Blue, Oregon Green 488, Oregon Green 514, Pacific Blue, rhodamine 6G, rhodamine green, rhodamine red, tetramethylrhodamine, Texas Red (available from Molecular Probes, Inc., Eugene, OR, USA).

The polypeptides of the present invention can also be conjugated to fluorophores, other proteins, and other macromolecules, using bifunctional linking reagents. Common homobifunctional reagents include, *e.g.*, APG, AEDP, BASED, BMB, BMDB, BMH, BMOE, BM[PEO]3, BM[PEO]4, BS3, BSOCOES, DFDNB, DMA, DMP, DMS, DPDPB, DSG, DSP (Lomant's Reagent), DSS, DST, DTBP, DTME, DTSSP, EGS, HBVS, Sulfo-BSOCOES, Sulfo-DST, Sulfo-EGS (all available from Pierce, Rockford, IL, USA); common heterobifunctional cross-linkers include ABH, AMAS, ANB-NOS, APDP, ASBA, BMPA, BMPH, BMPS, EDC, EMCA, EMCH, EMCS, KMUA, KMUH, GMBS, LC-SMCC, LC-SPDP, MBS, M2C2H, MPBH, MSA, NHS-ASA, PDPH, PMPI, SADP, SAED, SAND, SANPAH, SASD, SATP, SBAP, SFAD, SIA, SIAB, SMCC, SMPB, SMPH, SMPT, SPDP, Sulfo-EMCS, Sulfo-GMBS, Sulfo-HSAB, Sulfo-KMUS, Sulfo-LC-SPDP, Sulfo-MBS, Sulfo-NHS-LC-ASA, Sulfo-SADP, Sulfo-SANPAH, Sulfo-SIAB, Sulfo-SMCC, Sulfo-SMPB, Sulfo-LC-SMPT, SVSB, TFCS (all available Pierce, Rockford, IL, USA).

The polypeptides, fragments, and fusion proteins of the present invention can be conjugated, using such cross-linking reagents, to fluorophores that are not amine- or thiol-reactive. Other labels that usefully can be conjugated to the polypeptides, fragments, and fusion proteins of the present invention include radioactive labels, echosonographic contrast reagents, and MRI contrast agents.

The polypeptides, fragments, and fusion proteins of the present invention can also usefully be conjugated using cross-linking agents to carrier proteins, such as KLH, bovine thyroglobulin, and even bovine serum albumin (BSA), to increase immunogenicity for raising anti-CSP antibodies.

The polypeptides, fragments, and fusion proteins of the present invention can also usefully be conjugated to polyethylene glycol (PEG); PEGylation increases the serum half-life of proteins administered intravenously for replacement therapy. Delgado *et al.*, *Crit. Rev. Ther. Drug Carrier Syst.* 9(3-4): 249-304 (1992); Scott *et al.*, *Curr. Pharm. Des.* 4(6): 423-38 (1998); DeSantis *et al.*, *Curr. Opin. Biotechnol.* 10(4): 324-30 (1999), incorporated herein by reference in their entireties. PEG monomers can be attached to the protein directly or through a linker, with PEGylation using PEG monomers activated with tresyl chloride (2,2,2-trifluoroethanesulphonyl chloride) permitting direct attachment under mild conditions.

In yet another embodiment, the invention provides analogs of a polypeptide encoded by a nucleic acid molecule according to the instant invention. In a preferred embodiment, the polypeptide is a CSP. In a more preferred embodiment, the analog is derived from a polypeptide having part or all of the amino acid sequence of SEQ ID NO: 101 through 176. In a preferred embodiment, the analog is one that comprises one or more substitutions of non-natural amino acids or non-native inter-residue bonds compared to the naturally-occurring polypeptide. In general, the non-peptide analog is structurally similar to a CSP, but one or more peptide linkages is replaced by a linkage selected from the group consisting of --CH_2NH--, --CH_2S--, --CH_2-CH_2--, --CH=CH--(cis and trans), --$COCH_2$--, --$CH(OH)CH_2$-- and –CH_2SO--. In another embodiment, the non-peptide analog comprises substitution of one or more amino acids of a CSP with a D-amino acid of the same type or other non-natural amino acid in order to generate more stable peptides. D-amino acids can readily be incorporated during chemical peptide synthesis: peptides assembled from D-amino acids are more resistant to proteolytic attack; incorporation of D-amino acids can also be used to confer specific three-dimensional conformations on the peptide. Other amino acid analogues commonly added during chemical synthesis include ornithine, norleucine, phosphorylated amino acids (typically phosphoserine, phosphothreonine, phosphotyrosine), L-malonyltyrosine, a non-hydrolyzable analog of phosphotyrosine (*see, e.g.*, Kole *et al.*, *Biochem. Biophys. Res. Com.* 209: 817-821 (1995)), and various halogenated phenylalanine derivatives.

Non-natural amino acids can be incorporated during solid phase chemical synthesis or by recombinant techniques, although the former is typically more common. Solid phase chemical synthesis of peptides is well established in the art. Procedures are

described, inter alia, in Chan *et al.* (eds.), Fmoc Solid Phase Peptide Synthesis: A Practical Approach (Practical Approach Series), Oxford Univ. Press (March 2000); Jones, Amino Acid and Peptide Synthesis (Oxford Chemistry Primers, No 7), Oxford Univ. Press (1992); and Bodanszky, Principles of Peptide Synthesis (Springer Laboratory), Springer Verlag (1993); the disclosures of which are incorporated herein by reference in their entireties.

Amino acid analogues having detectable labels are also usefully incorporated during synthesis to provide derivatives and analogs. Biotin, for example can be added using biotinoyl-(9-fluorenylmethoxycarbonyl)-L-lysine (FMOC biocytin) (Molecular Probes, Eugene, OR, USA). Biotin can also be added enzymatically by incorporation into a fusion protein of a *E. coli* BirA substrate peptide. The FMOC and *t*BOC derivatives of dabcyl-L-lysine (Molecular Probes, Inc., Eugene, OR, USA) can be used to incorporate the dabcyl chromophore at selected sites in the peptide sequence during synthesis. The aminonaphthalene derivative EDANS, the most common fluorophore for pairing with the dabcyl quencher in fluorescence resonance energy transfer (FRET) systems, can be introduced during automated synthesis of peptides by using EDANS-FMOC-L-glutamic acid or the corresponding *t*BOC derivative (both from Molecular Probes, Inc., Eugene, OR, USA). Tetramethylrhodamine fluorophores can be incorporated during automated FMOC synthesis of peptides using (FMOC)-TMR-L-lysine (Molecular Probes, Inc. Eugene, OR, USA).

Other useful amino acid analogues that can be incorporated during chemical synthesis include aspartic acid, glutamic acid, lysine, and tyrosine analogues having allyl side-chain protection (Applied Biosystems, Inc., Foster City, CA, USA); the allyl side chain permits synthesis of cyclic, branched-chain, sulfonated, glycosylated, and phosphorylated peptides.

A large number of other FMOC-protected non-natural amino acid analogues capable of incorporation during chemical synthesis are available commercially, including, *e.g.*, Fmoc-2-aminobicyclo[2.2.1]heptane-2-carboxylic acid, Fmoc-3-endo-aminobicyclo[2.2.1]heptane-2-endo-carboxylic acid, Fmoc-3-exo-aminobicyclo[2.2.1]heptane-2-exo-carboxylic acid, Fmoc-3-endo-amino-bicyclo[2.2.1]hept-5-ene-2-endo-carboxylic acid, Fmoc-3-exo-amino-bicyclo[2.2.1]hept-5-ene-2-exo-carboxylic acid, Fmoc-cis-2-amino-1-cyclohexanecarboxylic acid, Fmoc-

trans-2-amino-1-cyclohexanecarboxylic acid, Fmoc-1-amino-1-cyclopentanecarboxylic acid, Fmoc-cis-2-amino-1-cyclopentanecarboxylic acid, Fmoc-1-amino-1-cyclopropanecarboxylic acid, Fmoc-D-2-amino-4-(ethylthio)butyric acid, Fmoc-L-2-amino-4-(ethylthio)butyric acid, Fmoc-L-buthionine, Fmoc-S-methyl-L-Cysteine, Fmoc-2-aminobenzoic acid (anthranillic acid), Fmoc-3-aminobenzoic acid, Fmoc-4-aminobenzoic acid, Fmoc-2-aminobenzophenone-2'-carboxylic acid, Fmoc-N-(4-aminobenzoyl)-β-alanine, Fmoc-2-amino-4,5-dimethoxybenzoic acid, Fmoc-4-aminohippuric acid, Fmoc-2-amino-3-hydroxybenzoic acid, Fmoc-2-amino-5-hydroxybenzoic acid, Fmoc-3-amino-4-hydroxybenzoic acid, Fmoc-4-amino-3-hydroxybenzoic acid, Fmoc-4-amino-2-hydroxybenzoic acid, Fmoc-5-amino-2-hydroxybenzoic acid, Fmoc-2-amino-3-methoxybenzoic acid, Fmoc-4-amino-3-methoxybenzoic acid, Fmoc-2-amino-3-methylbenzoic acid, Fmoc-2-amino-5-methylbenzoic acid, Fmoc-2-amino-6-methylbenzoic acid, Fmoc-3-amino-2-methylbenzoic acid, Fmoc-3-amino-4-methylbenzoic acid, Fmoc-4-amino-3-methylbenzoic acid, Fmoc-3-amino-2-naphtoic acid, Fmoc-D,L-3-amino-3-phenylpropionic acid, Fmoc-L-Methyldopa, Fmoc-2-amino-4,6-dimethyl-3-pyridinecarboxylic acid, Fmoc-D,L-amino-2-thiophenacetic acid, Fmoc-4-(carboxymethyl)piperazine, Fmoc-4-carboxypiperazine, Fmoc-4-(carboxymethyl)homopiperazine, Fmoc-4-phenyl-4-piperidinecarboxylic acid, Fmoc-L-1,2,3,4-tetrahydronorharman-3-carboxylic acid, Fmoc-L-thiazolidine-4-carboxylic acid, all available from The Peptide Laboratory (Richmond, CA, USA).

Non-natural residues can also be added biosynthetically by engineering a suppressor tRNA, typically one that recognizes the UAG stop codon, by chemical aminoacylation with the desired unnatural amino acid. Conventional site-directed mutagenesis is used to introduce the chosen stop codon UAG at the site of interest in the protein gene. When the acylated suppressor tRNA and the mutant gene are combined in an *in vitro* transcription/translation system, the unnatural amino acid is incorporated in response to the UAG codon to give a protein containing that amino acid at the specified position. Liu *et al.*, *Proc. Natl Acad. Sci. USA* 96(9): 4780-5 (1999); Wang *et al.*, *Science* 292(5516): 498-500 (2001).

Fusion Proteins

The present invention further provides fusions of each of the polypeptides and fragments of the present invention to heterologous polypeptides. In a preferred embodiment, the polypeptide is a CSP. In a more preferred embodiment, the polypeptide that is fused to the heterologous polypeptide comprises part or all of the amino acid sequence of SEQ ID NO: 101 through 176, or is a mutein, homologous polypeptide, analog or derivative thereof. In an even more preferred embodiment, the nucleic acid molecule encoding the fusion protein comprises all or part of the nucleic acid sequence of SEQ ID NO: 1 through 100, or comprises all or part of a nucleic acid sequence that selectively hybridizes or is homologous to a nucleic acid molecule comprising a nucleic acid sequence of SEQ ID NO: 1 through 100.

The fusion proteins of the present invention will include at least one fragment of the protein of the present invention, which fragment is at least 6, typically at least 8, often at least 15, and usefully at least 16, 17, 18, 19, or 20 amino acids long. The fragment of the protein of the present to be included in the fusion can usefully be at least 25 amino acids long, at least 50 amino acids long, and can be at least 75, 100, or even 150 amino acids long. Fusions that include the entirety of the proteins of the present invention have particular utility.

The heterologous polypeptide included within the fusion protein of the present invention is at least 6 amino acids in length, often at least 8 amino acids in length, and usefully at least 15, 20, and 25 amino acids in length. Fusions that include larger polypeptides, such as the IgG Fc region, and even entire proteins (such as GFP chromophore-containing proteins) are particular useful.

As described above in the description of vectors and expression vectors of the present invention, which discussion is incorporated here by reference in its entirety, heterologous polypeptides to be included in the fusion proteins of the present invention can usefully include those designed to facilitate purification and/or visualization of recombinantly-expressed proteins. *See, e.g.*, Ausubel, Chapter 16, (1992), *supra*. Although purification tags can also be incorporated into fusions that are chemically synthesized, chemical synthesis typically provides sufficient purity that further purification by HPLC suffices; however, visualization tags as above described retain their utility even when the protein is produced by chemical synthesis, and when so

included render the fusion proteins of the present invention useful as directly detectable markers of the presence of a polypeptide of the invention.

As also discussed above, heterologous polypeptides to be included in the fusion proteins of the present invention can usefully include those that facilitate secretion of recombinantly expressed proteins — into the periplasmic space or extracellular milieu for prokaryotic hosts, into the culture medium for eukaryotic cells — through incorporation of secretion signals and/or leader sequences. For example, a His^6 tagged protein can be purified on a Ni affinity column and a GST fusion protein can be purified on a glutathione affinity column. Similarly, a fusion protein comprising the Fc domain of IgG can be purified on a Protein A or Protein G column and a fusion protein comprising an epitope tag such as myc can be purified using an immunoaffinity column containing an anti-c-myc antibody. It is preferable that the epitope tag be separated from the protein encoded by the essential gene by an enzymatic cleavage site that can be cleaved after purification. See also the discussion of nucleic acid molecules encoding fusion proteins that may be expressed on the surface of a cell.

Other useful protein fusions of the present invention include those that permit use of the protein of the present invention as bait in a yeast two-hybrid system. *See* Bartel *et al.* (eds.), The Yeast Two-Hybrid System, Oxford University Press (1997); Zhu *et al.*, Yeast Hybrid Technologies, Eaton Publishing (2000); Fields *et al.*, *Trends Genet.* 10(8): 286-92 (1994); Mendelsohn *et al.*, *Curr. Opin. Biotechnol.* 5(5): 482-6 (1994); Luban *et al.*, *Curr. Opin. Biotechnol.* 6(1): 59-64 (1995); Allen *et al.*, *Trends Biochem. Sci.* 20(12): 511-6 (1995); Drees, *Curr. Opin. Chem. Biol.* 3(1): 64-70 (1999); Topcu *et al.*, *Pharm. Res.* 17(9): 1049-55 (2000); Fashena *et al.*, *Gene* 250(1-2): 1-14 (2000); ; Colas *et al.*, (1996) Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature* 380, 548-550; Norman, T. *et al.*, (1999) Genetic selection of peptide inhibitors of biological pathways. *Science* 285, 591-595, Fabbrizio *et al.*, (1999) Inhibition of mammalian cell proliferation by genetically selected peptide aptamers that functionally antagonize E2F activity. *Oncogene* 18, 4357-4363; Xu *et al.*, (1997) Cells that register logical relationships among proteins. *Proc Natl Acad Sci U S A.* 94, 12473-12478; Yang, *et al.*, (1995) Protein-peptide interactions analyzed with the yeast two-hybrid system. *Nuc. Acids Res.* 23, 1152-1156; Kolonin *et al.*, (1998) Targeting cyclin-dependent kinases in Drosophila with peptide aptamers. *Proc Natl Acad Sci U S A* 95,

14266-14271; Cohen *et al.*, (1998) An artificial cell-cycle inhibitor isolated from a combinatorial library. *Proc Natl Acad Sci U S A* 95, 14272-14277; Uetz, P.; Giot, L.; al, e.; Fields, S.; Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403, 623-627; Ito, *et al.*, (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98, 4569-4574, the disclosures of which are incorporated herein by reference in their entireties. Typically, such fusion is to either *E. coli* LexA or yeast GAL4 DNA binding domains. Related bait plasmids are available that express the bait fused to a nuclear localization signal.

Other useful fusion proteins include those that permit display of the encoded protein on the surface of a phage or cell, fusions to intrinsically fluorescent proteins, such as green fluorescent protein (GFP), and fusions to the IgG Fc region, as described above, which discussion is incorporated here by reference in its entirety.

The polypeptides and fragments of the present invention can also usefully be fused to protein toxins, such as *Pseudomonas* exotoxin A, *diphtheria* toxin, *shiga* toxin A, *anthrax* toxin lethal factor, ricin, in order to effect ablation of cells that bind or take up the proteins of the present invention.

Fusion partners include, *inter alia*, *myc*, hemagglutinin (HA), GST, immunoglobulins, β-galactosidase, biotin trpE, protein A, β-lactamase, -amylase, maltose binding protein, alcohol dehydrogenase, polyhistidine (for example, six histidine at the amino and/or carboxyl terminus of the polypeptide), lacZ, green fluorescent protein (GFP), yeast _ mating factor, GAL4 transcription activation or DNA binding domain, luciferase, and serum proteins such as ovalbumin, albumin and the constant domain of IgG. *See, e.g.*, Ausubel (1992), *supra* and Ausubel (1999), *supra*. Fusion proteins may also contain sites for specific enzymatic cleavage, such as a site that is recognized by enzymes such as Factor XIII, trypsin, pepsin, or any other enzyme known in the art. Fusion proteins will typically be made by either recombinant nucleic acid methods, as described above, chemically synthesized using techniques well-known in the art (*e.g.*, a Merrifield synthesis), or produced by chemical cross-linking.

Another advantage of fusion proteins is that the epitope tag can be used to bind the fusion protein to a plate or column through an affinity linkage for screening binding proteins or other molecules that bind to the CSP.

As further described below, the isolated polypeptides, muteins, fusion proteins, homologous proteins or allelic variants of the present invention can readily be used as specific immunogens to raise antibodies that specifically recognize CSPs, their allelic variants and homologues. The antibodies, in turn, can be used, *inter alia*, specifically to assay for the polypeptides of the present invention, particularly CSPs, *e.g.* by ELISA for detection of protein fluid samples, such as serum, by immunohistochemistry or laser scanning cytometry, for detection of protein in tissue samples, or by flow cytometry, for detection of intracellular protein in cell suspensions, for specific antibody-mediated isolation and/or purification of CSPs, as for example by immunoprecipitation, and for use as specific agonists or antagonists of CSPs.

One may determine whether polypeptides including muteins, fusion proteins, homologous proteins or allelic variants are functional by methods known in the art. For instance, residues that are tolerant of change while retaining function can be identified by altering the protein at known residues using methods known in the art, such as alanine scanning mutagenesis, Cunningham *et al.*, *Science* 244(4908): 1081-5 (1989); transposon linker scanning mutagenesis, Chen *et al.*, *Gene* 263(1-2): 39-48 (2001); combinations of homolog- and alanine-scanning mutagenesis, Jin *et al.*, *J. Mol. Biol.* 226(3): 851-65 (1992); combinatorial alanine scanning, Weiss *et al.*, *Proc. Natl. Acad. Sci USA* 97(16): 8950-4 (2000), followed by functional assay. Transposon linker scanning kits are available commercially (New England Biolabs, Beverly, MA, USA, catalog. no. E7-102S; EZ::TN™ In-Frame Linker Insertion Kit, catalogue no. EZI04KN, Epicentre Technologies Corporation, Madison, WI, USA).

Purification of the polypeptides including fragments, homologous polypeptides, muteins, analogs, derivatives and fusion proteins is well-known and within the skill of one having ordinary skill in the art. *See, e.g.,* Scopes, Protein Purification, 2d ed. (1987). Purification of recombinantly expressed polypeptides is described above. Purification of chemically-synthesized peptides can readily be effected, *e.g.*, by HPLC.

Accordingly, it is an aspect of the present invention to provide the isolated proteins of the present invention in pure or substantially pure form in the presence of absence of a stabilizing agent. Stabilizing agents include both proteinaceous or non-proteinaceous material and are well-known in the art. Stabilizing agents, such as albumin and polyethylene glycol (PEG) are known and are commercially available.

Although high levels of purity are preferred when the isolated proteins of the present invention are used as therapeutic agents, such as in vaccines and as replacement therapy, the isolated proteins of the present invention are also useful at lower purity. For example, partially purified proteins of the present invention can be used as immunogens to raise antibodies in laboratory animals.

In preferred embodiments, the purified and substantially purified proteins of the present invention are in compositions that lack detectable ampholytes, acrylamide monomers, bis-acrylamide monomers, and polyacrylamide.

The polypeptides, fragments, analogs, derivatives and fusions of the present invention can usefully be attached to a substrate. The substrate can be porous or solid, planar or non-planar; the bond can be covalent or noncovalent.

For example, the polypeptides, fragments, analogs, derivatives and fusions of the present invention can usefully be bound to a porous substrate, commonly a membrane, typically comprising nitrocellulose, polyvinylidene fluoride (PVDF), or cationically derivatized, hydrophilic PVDF; so bound, the proteins, fragments, and fusions of the present invention can be used to detect and quantify antibodies, *e.g.* in serum, that bind specifically to the immobilized protein of the present invention.

As another example, the polypeptides, fragments, analogs, derivatives and fusions of the present invention can usefully be bound to a substantially nonporous substrate, such as plastic, to detect and quantify antibodies, *e.g.* in serum, that bind specifically to the immobilized protein of the present invention. Such plastics include polymethylacrylic, polyethylene, polypropylene, polyacrylate, polymethylmethacrylate, polyvinylchloride, polytetrafluoroethylene, polystyrene, polycarbonate, polyacetal, polysulfone, celluloseacetate, cellulosenitrate, nitrocellulose, or mixtures thereof; when the assay is performed in a standard microtiter dish, the plastic is typically polystyrene.

The polypeptides, fragments, analogs, derivatives and fusions of the present invention can also be attached to a substrate suitable for use as a surface enhanced laser desorption ionization source; so attached, the protein, fragment, or fusion of the present invention is useful for binding and then detecting secondary proteins that bind with sufficient affinity or avidity to the surface-bound protein to indicate biologic interaction there between. The proteins, fragments, and fusions of the present invention can also be attached to a substrate suitable for use in surface plasmon resonance detection; so

attached, the protein, fragment, or fusion of the present invention is useful for binding and then detecting secondary proteins that bind with sufficient affinity or avidity to the surface-bound protein to indicate biological interaction there between.

Antibodies

In another aspect, the invention provides antibodies, including fragments and derivatives thereof, that bind specifically to polypeptides encoded by the nucleic acid molecules of the invention, as well as antibodies that bind to fragments, muteins, derivatives and analogs of the polypeptides. In a preferred embodiment, the antibodies are specific for a polypeptide that is a CSP, or a fragment, mutein, derivative, analog or fusion protein thereof. In a more preferred embodiment, the antibodies are specific for a polypeptide that comprises SEQ ID NO: 101 through 176, or a fragment, mutein, derivative, analog or fusion protein thereof.

The antibodies of the present invention can be specific for linear epitopes, discontinuous epitopes, or conformational epitopes of such proteins or protein fragments, either as present on the protein in its native conformation or, in some cases, as present on the proteins as denatured, as, *e.g.*, by solubilization in SDS. New epitopes may be also due to a difference in post translational modifications (PTMs) in disease versus normal tissue. For example, a particular site on a CSP may be glycosylated in cancerous cells, but not glycosylated in normal cells or visa versa. In addition, alternative splice forms of a CSP may be indicative of cancer. Differential degradation of the C or N-terminus of a CSP may also be a marker or target for anticancer therapy. For example, a CSP may be N-terminal degraded in cancer cells exposing new epitopes to which antibodies may selectively bind for diagnostic or therapeutic uses.

As is well-known in the art, the degree to which an antibody can discriminate as among molecular species in a mixture will depend, in part, upon the conformational relatedness of the species in the mixture; typically, the antibodies of the present invention will discriminate over adventitious binding to non-CSP polypeptides by at least 2-fold, more typically by at least 5-fold, typically by more than 10-fold, 25-fold, 50-fold, 75-fold, and often by more than 100-fold, and on occasion by more than 500-fold or 1000-fold. When used to detect the proteins or protein fragments of the present invention, the antibody of the present invention is sufficiently specific when it can be used to determine

the presence of the protein of the present invention in samples derived from human colon.

Typically, the affinity or avidity of an antibody (or antibody multimer, as in the case of an IgM pentamer) of the present invention for a protein or protein fragment of the present invention will be at least about 1×10^{-6} molar (M), typically at least about 5×10^{-7} M, 1×10^{-7} M, with affinities and avidities of at least 1×10^{-8} M, 5×10^{-9} M, 1×10^{-10} M and up to $1 \text{ X } 10^{-13}$ M proving especially useful.

The antibodies of the present invention can be naturally-occurring forms, such as IgG, IgM, IgD, IgE, IgY, and IgA, from any avian, reptilian, or mammalian species.

Human antibodies can, but will infrequently, be drawn directly from human donors or human cells. In this case, antibodies to the proteins of the present invention will typically have resulted from fortuitous immunization, such as autoimmune immunization, with the protein or protein fragments of the present invention. Such antibodies will typically, but will not invariably, be polyclonal. In addition, individual polyclonal antibodies may be isolated and cloned to generate monoclonals.

Human antibodies are more frequently obtained using transgenic animals that express human immunoglobulin genes, which transgenic animals can be affirmatively immunized with the protein immunogen of the present invention. Human Ig-transgenic mice capable of producing human antibodies and methods of producing human antibodies therefrom upon specific immunization are described, *inter alia*, in U.S. Patents 6,162,963; 6,150,584; 6,114,598; 6,075,181; 5,939,598; 5,877,397; 5,874,299; 5,814,318; 5,789,650; 5,770,429; 5,661,016; 5,633,425; 5,625,126; 5,569,825; 5,545,807; 5,545,806, and 5,591,669, the disclosures of which are incorporated herein by reference in their entireties. Such antibodies are typically monoclonal, and are typically produced using techniques developed for production of murine antibodies.

Human antibodies are particularly useful, and often preferred, when the antibodies of the present invention are to be administered to human beings as *in vivo* diagnostic or therapeutic agents, since recipient immune response to the administered antibody will often be substantially less than that occasioned by administration of an antibody derived from another species, such as mouse.

IgG, IgM, IgD, IgE, IgY, and IgA antibodies of the present invention can also be obtained from other species, including mammals such as rodents (typically mouse, but

also rat, guinea pig, and hamster) lagomorphs, typically rabbits, and also larger mammals, such as sheep, goats, cows, and horses, and other egg laying birds or reptiles such as chickens or alligators. For example, avian antibodies may be generated using techniques described in WO 00/29444, published 25 May 2000, the contents of which are hereby incorporated in their entirety. In such cases, as with the transgenic human-antibody-producing non-human mammals, fortuitous immunization is not required, and the non-human mammal is typically affirmatively immunized, according to standard immunization protocols, with the protein or protein fragment of the present invention.

As discussed above, virtually all fragments of 8 or more contiguous amino acids of the proteins of the present invention can be used effectively as immunogens when conjugated to a carrier, typically a protein such as bovine thyroglobulin, keyhole limpet hemocyanin, or bovine serum albumin, conveniently using a bifunctional linker such as those described elsewhere above, which discussion is incorporated by reference here.

Immunogenicity can also be conferred by fusion of the polypeptide and fragments of the present invention to other moieties. For example, peptides of the present invention can be produced by solid phase synthesis on a branched polylysine core matrix; these multiple antigenic peptides (MAPs) provide high purity, increased avidity, accurate chemical definition and improved safety in vaccine development. Tam *et al.*, *Proc. Natl. Acad. Sci. USA* 85: 5409-5413 (1988); Posnett *et al.*, *J. Biol. Chem.* 263: 1719-1725 (1988).

Protocols for immunizing non-human mammals or avian species are well-established in the art. *See* Harlow *et al.* (eds.), Using Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory (1998); Coligan *et al.* (eds.), Current Protocols in Immunology, John Wiley & Sons, Inc. (2001); Zola, Monoclonal Antibodies: Preparation and Use of Monoclonal Antibodies and Engineered Antibody Derivatives (Basics: From Background to Bench), Springer Verlag (2000); Gross M, Speck *J.Dtsch. Tierarztl. Wochenschr.* 103: 417-422 (1996), the disclosures of which are incorporated herein by reference. Immunization protocols often include multiple immunizations, either with or without adjuvants such as Freund's complete adjuvant and Freund's incomplete adjuvant, and may include naked DNA immunization (Moss, *Semin. Immunol.* 2: 317-327 (1990).

Antibodies from non-human mammals and avian species can be polyclonal or monoclonal, with polyclonal antibodies having certain advantages in

immunohistochemical detection of the proteins of the present invention and monoclonal antibodies having advantages in identifying and distinguishing particular epitopes of the proteins of the present invention. Antibodies from avian species may have particular advantage in detection of the proteins of the present invention, in human serum or tissues (Vikinge et al., *Biosens. Bioelectron.* 13: 1257-1262 (1998).

Following immunization, the antibodies of the present invention can be produced using any art-accepted technique. Such techniques are well-known in the art, Coligan, *supra*; Zola, *supra*; Howard *et al.* (eds.), Basic Methods in Antibody Production and Characterization, CRC Press (2000); Harlow, *supra*; Davis (ed.), Monoclonal Antibody Protocols, Vol. 45, Humana Press (1995); Delves (ed.), Antibody Production: Essential Techniques, John Wiley & Son Ltd (1997); Kenney, Antibody Solution: An Antibody Methods Manual, Chapman & Hall (1997), incorporated herein by reference in their entireties, and thus need not be detailed here.

Briefly, however, such techniques include, *inter alia*, production of monoclonal antibodies by hybridomas and expression of antibodies or fragments or derivatives thereof from host cells engineered to express immunoglobulin genes or fragments thereof. These two methods of production are not mutually exclusive: genes encoding antibodies specific for the proteins or protein fragments of the present invention can be cloned from hybridomas and thereafter expressed in other host cells. Nor need the two necessarily be performed together: *e.g.*, genes encoding antibodies specific for the proteins and protein fragments of the present invention can be cloned directly from B cells known to be specific for the desired protein, as further described in U.S Patent 5,627,052, the disclosure of which is incorporated herein by reference in its entirety, or from antibody-displaying phage.

Recombinant expression in host cells is particularly useful when fragments or derivatives of the antibodies of the present invention are desired.

Host cells for recombinant production of either whole antibodies, antibody fragments, or antibody derivatives can be prokaryotic or eukaryotic.

Prokaryotic hosts are particularly useful for producing phage displayed antibodies of the present invention.

The technology of phage-displayed antibodies, in which antibody variable region fragments are fused, for example, to the gene III protein (pIII) or gene VIII protein

(pVIII) for display on the surface of filamentous phage, such as M13, is by now well-established. *See, e.g.,* Sidhu, *Curr. Opin. Biotechnol.* 11(6): 610-6 (2000); Griffiths *et al., Curr. Opin. Biotechnol.* 9(1): 102-8 (1998); Hoogenboom *et al., Immunotechnology,* 4(1): 1-20 (1998); Rader *et al., Current Opinion in Biotechnology* 8: 503-508 (1997); Aujame *et al., Human Antibodies* 8: 155-168 (1997); Hoogenboom, *Trends in Biotechnol.* 15: 62-70 (1997); de Kruif *et al.,* 17: 453-455 (1996); Barbas *et al., Trends in Biotechnol.* 14: 230-234 (1996); Winter *et al., Ann. Rev. Immunol.* 433-455 (1994). Techniques and protocols required to generate, propagate, screen (pan), and use the antibody fragments from such libraries have recently been compiled. *See, e.g.,* Barbas (2001), *supra*; Kay, *supra*; Abelson, *supra*, the disclosures of which are incorporated herein by reference in their entireties.

Typically, phage-displayed antibody fragments are scFv fragments or Fab fragments; when desired, full length antibodies can be produced by cloning the variable regions from the displaying phage into a complete antibody and expressing the full length antibody in a further prokaryotic or a eukaryotic host cell.

Eukaryotic cells are also useful for expression of the antibodies, antibody fragments, and antibody derivatives of the present invention.

For example, antibody fragments of the present invention can be produced in *Pichia pastoris* and in *Saccharomyces cerevisiae. See, e.g.,* Takahashi *et al., Biosci. Biotechnol. Biochem.* 64(10): 2138-44 (2000); Freyre *et al.,* J. Biotechnol. 76(2-3):1 57-63 (2000); Fischer *et al., Biotechnol. Appl. Biochem.* 30 (Pt 2): 117-20 (1999); Pennell *et al., Res. Immunol.* 149(6): 599-603 (1998); Eldin *et al., J. Immunol. Methods.* 201(1): 67-75 (1997);, Frenken *et al., Res. Immunol.* 149(6): 589-99 (1998); Shusta *et al., Nature Biotechnol.* 16(8): 773-7 (1998), the disclosures of which are incorporated herein by reference in their entireties.

Antibodies, including antibody fragments and derivatives, of the present invention can also be produced in insect cells. *See, e.g.,* Li *et al., Protein Expr. Purif.* 21(1): 121-8 (2001); Ailor *et al., Biotechnol. Bioeng.* 58(2-3): 196-203 (1998); Hsu *et al., Biotechnol. Prog.* 13(1): 96-104 (1997); Edelman *et al., Immunology* 91(1): 13-9 (1997); and Nesbit *et al., J. Immunol. Methods* 151(1-2): 201-8 (1992), the disclosures of which are incorporated herein by reference in their entireties.

Antibodies and fragments and derivatives thereof of the present invention can also be produced in plant cells, particularly maize or tobacco, Giddings *et al.*, *Nature Biotechnol.* 18(11): 1151-5 (2000); Gavilondo *et al.*, *Biotechniques* 29(1): 128-38 (2000); Fischer *et al.*, *J. Biol. Regul. Homeost. Agents* 14(2): 83-92 (2000); Fischer *et al.*, *Biotechnol. Appl. Biochem.* 30 (Pt 2): 113-6 (1999); Fischer *et al.*, *Biol. Chem.* 380(7-8): 825-39 (1999); Russell, *Curr. Top. Microbiol. Immunol.* 240: 119-38 (1999); and Ma *et al.*, *Plant Physiol.* 109(2): 341-6 (1995), the disclosures of which are incorporated herein by reference in their entireties.

Antibodies, including antibody fragments and derivatives, of the present invention can also be produced in transgenic, non-human, mammalian milk. *See, e.g.* Pollock et al., *J. Immunol Methods.* 231: 147-57 (1999); Young et al., *Res. Immunol.* 149: 609-10 (1998); Limonta et al., *Immunotechnology* 1: 107-13 (1995), the disclosures of which are incorporated herein by reference in their entireties.

Mammalian cells useful for recombinant expression of antibodies, antibody fragments, and antibody derivatives of the present invention include CHO cells, COS cells, 293 cells, and myeloma cells.

Verma *et al.*, *J. Immunol. Methods* 216(1-2):165-81 (1998), herein incorporated by reference, review and compare bacterial, yeast, insect and mammalian expression systems for expression of antibodies.

Antibodies of the present invention can also be prepared by cell free translation, as further described in Merk *et al.*, *J. Biochem.* (Tokyo) 125(2): 328-33 (1999) and Ryabova *et al.*, *Nature Biotechnol.* 15(1): 79-84 (1997), and in the milk of transgenic animals, as further described in Pollock *et al.*, *J. Immunol. Methods* 231(1-2): 147-57 (1999), the disclosures of which are incorporated herein by reference in their entireties.

The invention further provides antibody fragments that bind specifically to one or more of the proteins and protein fragments of the present invention, to one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention, or the binding of which can be competitively inhibited by one or more of the proteins and protein fragments of the present invention or one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention.

Among such useful fragments are Fab, Fab', Fv, F(ab)'$_2$, and single chain Fv (scFv) fragments. Other useful fragments are described in Hudson, *Curr. Opin. Biotechnol.* 9(4): 395-402 (1998).

It is also an aspect of the present invention to provide antibody derivatives that bind specifically to one or more of the proteins and protein fragments of the present invention, to one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention, or the binding of which can be competitively inhibited by one or more of the proteins and protein fragments of the present invention or one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention.

Among such useful derivatives are chimeric, primatized, and humanized antibodies; such derivatives are less immunogenic in human beings, and thus more suitable for *in vivo* administration, than are unmodified antibodies from non-human mammalian species. Another useful derivative is PEGylation to increase the serum half life of the antibodies.

Chimeric antibodies typically include heavy and/or light chain variable regions (including both CDR and framework residues) of immunoglobulins of one species, typically mouse, fused to constant regions of another species, typically human. *See, e.g.*, United States Patent No. 5,807,715; Morrison *et al.*, *Proc. Natl. Acad. Sci USA*.81(21): 6851-5 (1984); Sharon *et al.*, *Nature* 309(5966): 364-7 (1984); Takeda *et al.*, *Nature* 314(6010): 452-4 (1985), the disclosures of which are incorporated herein by reference in their entireties. Primatized and humanized antibodies typically include heavy and/or light chain CDRs from a murine antibody grafted into a non-human primate or human antibody V region framework, usually further comprising a human constant region, Riechmann *et al.*, *Nature* 332(6162): 323-7 (1988); Co *et al.*, *Nature* 351(6326): 501-2 (1991); United States Patent Nos. 6,054,297; 5,821,337; 5,770,196; 5,766,886; 5,821,123; 5,869,619; 6,180,377; 6,013,256; 5,693,761; and 6,180,370, the disclosures of which are incorporated herein by reference in their entireties.

Other useful antibody derivatives of the invention include heteromeric antibody complexes and antibody fusions, such as diabodies (bispecific antibodies), single-chain diabodies, and intrabodies.

It is contemplated that the nucleic acids encoding the antibodies of the present invention can be operably joined to other nucleic acids forming a recombinant vector for cloning or for expression of the antibodies of the invention. The present invention includes any recombinant vector containing the coding sequences, or part thereof, whether for eukaryotic transduction, transfection or gene therapy. Such vectors may be prepared using conventional molecular biology techniques, known to those with skill in the art, and would comprise DNA encoding sequences for the immunoglobulin V-regions including framework and CDRs or parts thereof, and a suitable promoter either with or without a signal sequence for intracellular transport. Such vectors may be transduced or transfected into eukaryotic cells or used for gene therapy (Marasco et al., *Proc. Natl. Acad. Sci. (USA)* 90: 7889-7893 (1993); Duan et al., *Proc. Natl. Acad. Sci. (USA)* 91: 5075-5079 (1994), by conventional techniques, known to those with skill in the art.

The antibodies of the present invention, including fragments and derivatives thereof, can usefully be labeled. It is, therefore, another aspect of the present invention to provide labeled antibodies that bind specifically to one or more of the proteins and protein fragments of the present invention, to one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention, or the binding of which can be competitively inhibited by one or more of the proteins and protein fragments of the present invention or one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention.

The choice of label depends, in part, upon the desired use.

For example, when the antibodies of the present invention are used for immunohistochemical staining of tissue samples, the label is preferably an enzyme that catalyzes production and local deposition of a detectable product.

Enzymes typically conjugated to antibodies to permit their immunohistochemical visualization are well-known, and include alkaline phosphatase, β-galactosidase, glucose oxidase, horseradish peroxidase (HRP), and urease. Typical substrates for production and deposition of visually detectable products include o-nitrophenyl-beta-D-galactopyranoside (ONPG); o-phenylenediamine dihydrochloride (OPD); p-nitrophenyl phosphate (PNPP); p-nitrophenyl-beta-D-galactopryanoside (PNPG); 3',3'-diaminobenzidine (DAB); 3-amino-9-ethylcarbazole (AEC); 4-chloro-1-naphthol (CN); 5-bromo-4-chloro-3-indolyl-phosphate (BCIP); ABTS®; BluoGal; iodonitrotetrazolium

(INT); nitroblue tetrazolium chloride (NBT); phenazine methosulfate (PMS); phenolphthalein monophosphate (PMP); tetramethyl benzidine (TMB); tetranitroblue tetrazolium (TNBT); X-Gal; X-Gluc; and X-Glucoside.

Other substrates can be used to produce products for local deposition that are luminescent. For example, in the presence of hydrogen peroxide (H_2O_2), horseradish peroxidase (HRP) can catalyze the oxidation of cyclic diacylhydrazides, such as luminol. Immediately following the oxidation, the luminol is in an excited state (intermediate reaction product), which decays to the ground state by emitting light. Strong enhancement of the light emission is produced by enhancers, such as phenolic compounds. Advantages include high sensitivity, high resolution, and rapid detection without radioactivity and requiring only small amounts of antibody. *See, e.g.*, Thorpe *et al.*, *Methods Enzymol.* 133: 331-53 (1986); Kricka *et al.*, *J. Immunoassay* 17(1): 67-83 (1996); and Lundqvist *et al.*, *J. Biolumin. Chemilumin.* 10(6): 353-9 (1995), the disclosures of which are incorporated herein by reference in their entireties. Kits for such enhanced chemiluminescent detection (ECL) are available commercially.

The antibodies can also be labeled using colloidal gold.

As another example, when the antibodies of the present invention are used, *e.g.*, for flow cytometric detection, for scanning laser cytometric detection, or for fluorescent immunoassay, they can usefully be labeled with fluorophores.

There are a wide variety of fluorophore labels that can usefully be attached to the antibodies of the present invention.

For flow cytometric applications, both for extracellular detection and for intracellular detection, common useful fluorophores can be fluorescein isothiocyanate (FITC), allophycocyanin (APC), R-phycoerythrin (PE), peridinin chlorophyll protein (PerCP), Texas Red, Cy3, Cy5, fluorescence resonance energy tandem fluorophores such as PerCP-Cy5.5, PE-Cy5, PE-Cy5.5, PE-Cy7, PE-Texas Red, and APC-Cy7.

Other fluorophores include, *inter alia*, Alexa Fluor® 350, Alexa Fluor® 488, Alexa Fluor® 532, Alexa Fluor® 546, Alexa Fluor® 568, Alexa Fluor® 594, Alexa Fluor® 647 (monoclonal antibody labeling kits available from Molecular Probes, Inc., Eugene, OR, USA), BODIPY dyes, such as BODIPY 493/503, BODIPY FL, BODIPY R6G, BODIPY 530/550, BODIPY TMR, BODIPY 558/568, BODIPY 558/568, BODIPY 564/570, BODIPY 576/589, BODIPY 581/591, BODIPY TR, BODIPY

630/650, BODIPY 650/665, Cascade Blue, Cascade Yellow, Dansyl, lissamine rhodamine B, Marina Blue, Oregon Green 488, Oregon Green 514, Pacific Blue, rhodamine 6G, rhodamine green, rhodamine red, tetramethylrhodamine, Texas Red (available from Molecular Probes, Inc., Eugene, OR, USA), and Cy2, Cy3, Cy3.5, Cy5, Cy5.5, Cy7, all of which are also useful for fluorescently labeling the antibodies of the present invention.

For secondary detection using labeled avidin, streptavidin, captavidin or neutravidin, the antibodies of the present invention can usefully be labeled with biotin.

When the antibodies of the present invention are used, *e.g.*, for Western blotting applications, they can usefully be labeled with radioisotopes, such as ^{33}P, ^{32}P, ^{35}S, ^{3}H, and ^{125}I.

As another example, when the antibodies of the present invention are used for radioimmunotherapy, the label can usefully be ^{228}Th, ^{227}Ac, ^{225}Ac, ^{223}Ra, ^{213}Bi, ^{212}Pb, ^{212}Bi, ^{211}At, ^{203}Pb, ^{194}Os, ^{188}Re, ^{186}Re, ^{153}Sm, ^{149}Tb, ^{131}I, ^{125}I, ^{111}In, ^{105}Rh, ^{99m}Tc, ^{97}Ru, ^{90}Y, ^{90}Sr, ^{88}Y, ^{72}Se, ^{67}Cu, or ^{47}Sc.

As another example, when the antibodies of the present invention are to be used for *in vivo* diagnostic use, they can be rendered detectable by conjugation to MRI contrast agents, such as gadolinium diethylenetriaminepentaacetic acid (DTPA), Lauffer *et al.*, *Radiology* 207(2): 529-38 (1998), or by radioisotopic labeling.

As would be understood, use of the labels described above is not restricted to the application for which they are mentioned.

The antibodies of the present invention, including fragments and derivatives thereof, can also be conjugated to toxins, in order to target the toxin's ablative action to cells that display and/or express the proteins of the present invention. Commonly, the antibody in such immunotoxins is conjugated to *Pseudomonas* exotoxin A, *diphtheria* toxin, *shiga* toxin A, *anthrax* toxin lethal factor, or ricin. *See* Hall (ed.), Immunotoxin Methods and Protocols (Methods in Molecular Biology, vol. 166), Humana Press (2000); and Frankel *et al.* (eds.), Clinical Applications of Immunotoxins, Springer-Verlag (1998), the disclosures of which are incorporated herein by reference in their entireties.

The antibodies of the present invention can usefully be attached to a substrate, and it is, therefore, another aspect of the invention to provide antibodies that bind specifically to one or more of the proteins and protein fragments of the present invention,

to one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention, or the binding of which can be competitively inhibited by one or more of the proteins and protein fragments of the present invention or one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention, attached to a substrate.

Substrates can be porous or nonporous, planar or nonplanar.

For example, the antibodies of the present invention can usefully be conjugated to filtration media, such as NHS-activated Sepharose or CNBr-activated Sepharose for purposes of immunoaffinity chromatography.

For example, the antibodies of the present invention can usefully be attached to paramagnetic microspheres, typically by biotin-streptavidin interaction, which microspheres can then be used for isolation of cells that express or display the proteins of the present invention. As another example, the antibodies of the present invention can usefully be attached to the surface of a microtiter plate for ELISA.

As noted above, the antibodies of the present invention can be produced in prokaryotic and eukaryotic cells. It is, therefore, another aspect of the present invention to provide cells that express the antibodies of the present invention, including hybridoma cells, B cells, plasma cells, and host cells recombinantly modified to express the antibodies of the present invention.

In yet a further aspect, the present invention provides aptamers evolved to bind specifically to one or more of the proteins and protein fragments of the present invention, to one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention, or the binding of which can be competitively inhibited by one or more of the proteins and protein fragments of the present invention or one or more of the proteins and protein fragments encoded by the isolated nucleic acids of the present invention.

In sum, one of skill in the art, provided with the teachings of this invention, has available a variety of methods which may be used to alter the biological properties of the antibodies of this invention including methods which would increase or decrease the stability or half-life, immunogenicity, toxicity, affinity or yield of a given antibody molecule, or to alter it in any other way that may render it more suitable for a particular application.

Transgenic Animals and Cells

In another aspect, the invention provides transgenic cells and non-human organisms comprising nucleic acid molecules of the invention. In a preferred embodiment, the transgenic cells and non-human organisms comprise a nucleic acid molecule encoding a CSP. In a preferred embodiment, the CSP comprises an amino acid sequence selected from SEQ ID NO: 101 through 176, or a fragment, mutein, homologous protein or allelic variant thereof. In another preferred embodiment, the transgenic cells and non-human organism comprise a CSNA of the invention, preferably a CSNA comprising a nucleotide sequence selected from the group consisting of SEQ ID NO: 1 through 100, or a part, substantially similar nucleic acid molecule, allelic variant or hybridizing nucleic acid molecule thereof.

In another embodiment, the transgenic cells and non-human organisms have a targeted disruption or replacement of the endogenous orthologue of the human CSG. The transgenic cells can be embryonic stem cells or somatic cells. The transgenic non-human organisms can be chimeric, nonchimeric heterozygotes, and nonchimeric homozygotes. Methods of producing transgenic animals are well-known in the art. *See, e.g.*, Hogan *et al.*, Manipulating the Mouse Embryo: A Laboratory Manual, 2d ed., Cold Spring Harbor Press (1999); Jackson *et al.*, Mouse Genetics and Transgenics: A Practical Approach, Oxford University Press (2000); and Pinkert, Transgenic Animal Technology: A Laboratory Handbook, Academic Press (1999).

Any technique known in the art may be used to introduce a nucleic acid molecule of the invention into an animal to produce the founder lines of transgenic animals. Such techniques include, but are not limited to, pronuclear microinjection. (*see, e.g.*, Paterson *et al., Appl. Microbiol. Biotechnol.* 40: 691-698 (1994); Carver *et al., Biotechnology* 11: 1263-1270 (1993); Wright *et al., Biotechnology* 9: 830-834 (1991); and U.S. Patent 4,873,191 (1989 retrovirus-mediated gene transfer into germ lines, blastocysts or embryos (*see, e.g.,* Van der Putten *et al., Proc. Natl. Acad. Sci., USA* 82: 6148-6152 (1985)); gene targeting in embryonic stem cells (*see, e.g.,* Thompson *et al., Cell* 56: 313-321 (1989)); electroporation of cells or embryos (*see, e.g.,* Lo, 1983, *Mol. Cell. Biol.* 3: 1803-1814 (1983)); introduction using a gene gun (*see, e.g.,* Ulmer *et al., Science* 259: 1745-49 (1993); introducing nucleic acid constructs into embryonic pleuripotent stem

cells and transferring the stem cells back into the blastocyst; and sperm-mediated gene transfer (*see, e.g.,* Lavitrano *et al., Cell* 57: 717-723 (1989)).

Other techniques include, for example, nuclear transfer into enucleated oocytes of nuclei from cultured embryonic, fetal, or adult cells induced to quiescence (*see, e.g.,* Campell *et al., Nature* 380: 64-66 (1996); Wilmut *et al., Nature* 385: 810-813 (1997)). The present invention provides for transgenic animals that carry the transgene (*i.e.*, a nucleic acid molecule of the invention) in all their cells, as well as animals which carry the transgene in some, but not all their cells, i. e., mosaic animals or chimeric animals.

The transgene may be integrated as a single transgene or as multiple copies, such as in concatamers, e. g., head-to-head tandems or head-to-tail tandems. The transgene may also be selectively introduced into and activated in a particular cell type by following, *e.g.*, the teaching of Lasko *et al. et al., Proc. Natl. Acad. Sci. USA* 89: 6232-6236 (1992). The regulatory sequences required for such a cell-type specific activation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art.

Once transgenic animals have been generated, the expression of the recombinant gene may be assayed utilizing standard techniques. Initial screening may be accomplished by Southern blot analysis or PCR techniques to analyze animal tissues to verify that integration of the transgene has taken place. The level of mRNA expression of the transgene in the tissues of the transgenic animals may also be assessed using techniques which include, but are not limited to, Northern blot analysis of tissue samples obtained from the animal, *in situ* hybridization analysis, and reverse transcriptase-PCR (RT-PCR). Samples of transgenic gene-expressing tissue may also be evaluated immunocytochemically or immunohistochemically using antibodies specific for the transgene product.

Once the founder animals are produced, they may be bred, inbred, outbred, or crossbred to produce colonies of the particular animal. Examples of such breeding strategies include, but are not limited to: outbreeding of founder animals with more than one integration site in order to establish separate lines; inbreeding of separate lines in order to produce compound transgenics that express the transgene at higher levels because of the effects of additive expression of each transgene; crossing of heterozygous transgenic animals to produce animals homozygous for a given integration site in order to

both augment expression and eliminate the need for screening of animals by DNA analysis; crossing of separate homozygous lines to produce compound heterozygous or homozygous lines; and breeding to place the transgene on a distinct background that is appropriate for an experimental model of interest.

Transgenic animals of the invention have uses which include, but are not limited to, animal model systems useful in elaborating the biological function of polypeptides of the present invention, studying conditions and/or disorders associated with aberrant expression, and in screening for compounds effective in ameliorating such conditions and/or disorders.

Methods for creating a transgenic animal with a disruption of a targeted gene are also well-known in the art. In general, a vector is designed to comprise some nucleotide sequences homologous to the endogenous targeted gene. The vector is introduced into a cell so that it may integrate, via homologous recombination with chromosomal sequences, into the endogenous gene, thereby disrupting the function of the endogenous gene. The transgene may also be selectively introduced into a particular cell type, thus inactivating the endogenous gene in only that cell type. *See, e.g.,* Gu *et al.*, *Science* 265: 103-106 (1994). The regulatory sequences required for such a cell-type specific inactivation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art. *See, e.g.*, Smithies *et al.*, *Nature* 317: 230-234 (1985); Thomas *et al.*, *Cell* 51: 503-512 (1987); Thompson *et al.*, *Cell* 5: 313-321 (1989).

In one embodiment, a mutant, non-functional nucleic acid molecule of the invention (or a completely unrelated DNA sequence) flanked by DNA homologous to the endogenous nucleic acid sequence (either the coding regions or regulatory regions of the gene) can be used, with or without a selectable marker and/or a negative selectable marker, to transfect cells that express polypeptides of the invention *in vivo*. In another embodiment, techniques known in the art are used to generate knockouts in cells that contain, but do not express the gene of interest. Insertion of the DNA construct, via targeted homologous recombination, results in inactivation of the targeted gene. Such approaches are particularly suited in research and agricultural fields where modifications to embryonic stem cells can be used to generate animal offspring with an inactive targeted gene. *See, e.g.,* Thomas, *supra* and Thompson, *supra*. However this approach can be routinely adapted for use in humans provided the recombinant DNA constructs are

directly administered or targeted to the required site *in vivo* using appropriate viral vectors that will be apparent to those of skill in the art.

In further embodiments of the invention, cells that are genetically engineered to express the polypeptides of the invention, or alternatively, that are genetically engineered not to express the polypeptides of the invention (*e.g.*, knockouts) are administered to a patient *in vivo*. Such cells may be obtained from an animal or patient or an MHC compatible donor and can include, but are not limited to fibroblasts, bone marrow cells, blood cells (*e.g.*, lymphocytes), adipocytes, muscle cells, endothelial cells etc. The cells are genetically engineered *in vitro* using recombinant DNA techniques to introduce the coding sequence of polypeptides of the invention into the cells, or alternatively, to disrupt the coding sequence and/or endogenous regulatory sequence associated with the polypeptides of the invention, *e.g.*, by transduction (using viral vectors, and preferably vectors that integrate the transgene into the cell genome) or transfection procedures, including, but not limited to, the use of plasmids, cosmids, YACs, naked DNA, electroporation, liposomes, etc.

The coding sequence of the polypeptides of the invention can be placed under the control of a strong constitutive or inducible promoter or promoter/enhancer to achieve expression, and preferably secretion, of the polypeptides of the invention. The engineered cells which express and preferably secrete the polypeptides of the invention can be introduced into the patient systemically, *e.g.*, in the circulation, or intraperitoneally.

Alternatively, the cells can be incorporated into a matrix and implanted in the body, *e.g.*, genetically engineered fibroblasts can be implanted as part of a skin graft; genetically engineered endothelial cells can be implanted as part of a lymphatic or vascular graft. *See, e.g.,* U.S. Patents 5,399,349 and 5,460,959, each of which is incorporated by reference herein in its entirety.

When the cells to be administered are non-autologous or non-MHC compatible cells, they can be administered using well-known techniques which prevent the development of a host immune response against the introduced cells. For example, the cells may be introduced in an encapsulated form which, while allowing for an exchange of components with the immediate extracellular environment, does not allow the introduced cells to be recognized by the host immune system.

Transgenic and "knock-out" animals of the invention have uses which include, but are not limited to, animal model systems useful in elaborating the biological function of polypeptides of the present invention, studying conditions and/or disorders associated with aberrant expression, and in screening for compounds effective in ameliorating such conditions and/or disorders.

Computer Readable Means

A further aspect of the invention relates to a computer readable means for storing the nucleic acid and amino acid sequences of the instant invention. In a preferred embodiment, the invention provides a computer readable means for storing SEQ ID NO: 1 through 100 and SEQ ID NO: 101 through 176 as described herein, as the complete set of sequences or in any combination. The records of the computer readable means can be accessed for reading and display and for interface with a computer system for the application of programs allowing for the location of data upon a query for data meeting certain criteria, the comparison of sequences, the alignment or ordering of sequences meeting a set of criteria, and the like.

The nucleic acid and amino acid sequences of the invention are particularly useful as components in databases useful for search analyses as well as in sequence analysis algorithms. As used herein, the terms "nucleic acid sequences of the invention" and "amino acid sequences of the invention" mean any detectable chemical or physical characteristic of a polynucleotide or polypeptide of the invention that is or may be reduced to or stored in a computer readable form. These include, without limitation, chromatographic scan data or peak data, photographic data or scan data therefrom, and mass spectrographic data.

This invention provides computer readable media having stored thereon sequences of the invention. A computer readable medium may comprise one or more of the following: a nucleic acid sequence comprising a sequence of a nucleic acid sequence of the invention; an amino acid sequence comprising an amino acid sequence of the invention; a set of nucleic acid sequences wherein at least one of said sequences comprises the sequence of a nucleic acid sequence of the invention; a set of amino acid sequences wherein at least one of said sequences comprises the sequence of an amino acid sequence of the invention; a data set representing a nucleic acid sequence comprising the sequence of one or more nucleic acid sequences of the invention; a data

set representing a nucleic acid sequence encoding an amino acid sequence comprising the sequence of an amino acid sequence of the invention; a set of nucleic acid sequences wherein at least one of said sequences comprises the sequence of a nucleic acid sequence of the invention; a set of amino acid sequences wherein at least one of said sequences comprises the sequence of an amino acid sequence of the invention; a data set representing a nucleic acid sequence comprising the sequence of a nucleic acid sequence of the invention; a data set representing a nucleic acid sequence encoding an amino acid sequence comprising the sequence of an amino acid sequence of the invention. The computer readable medium can be any composition of matter used to store information or data, including, for example, commercially available floppy disks, tapes, hard drives, compact disks, and video disks.

Also provided by the invention are methods for the analysis of character sequences, particularly genetic sequences. Preferred methods of sequence analysis include, for example, methods of sequence homology analysis, such as identity and similarity analysis, RNA structure analysis, sequence assembly, cladistic analysis, sequence motif analysis, open reading frame determination, nucleic acid base calling, and sequencing chromatogram peak analysis.

A computer-based method is provided for performing nucleic acid sequence identity or similarity identification. This method comprises the steps of providing a nucleic acid sequence comprising the sequence of a nucleic acid of the invention in a computer readable medium; and comparing said nucleic acid sequence to at least one nucleic acid or amino acid sequence to identify sequence identity or similarity.

A computer-based method is also provided for performing amino acid homology identification, said method comprising the steps of: providing an amino acid sequence comprising the sequence of an amino acid of the invention in a computer readable medium; and comparing said an amino acid sequence to at least one nucleic acid or an amino acid sequence to identify homology.

A computer-based method is still further provided for assembly of overlapping nucleic acid sequences into a single nucleic acid sequence, said method comprising the steps of: providing a first nucleic acid sequence comprising the sequence of a nucleic acid of the invention in a computer readable medium; and screening for at least one

overlapping region between said first nucleic acid sequence and a second nucleic acid sequence.

Diagnostic Methods for Colon Cancer

The present invention also relates to quantitative and qualitative diagnostic assays and methods for detecting, diagnosing, monitoring, staging and predicting cancers by comparing expression of a CSNA or a CSP in a human patient that has or may have colon cancer, or who is at risk of developing colon cancer, with the expression of a CSNA or a CSP in a normal human control. For purposes of the present invention, "expression of a CSNA" or "CSNA expression" means the quantity of CSG mRNA that can be measured by any method known in the art or the level of transcription that can be measured by any method known in the art in a cell, tissue, organ or whole patient. Similarly, the term "expression of a CSP" or "CSP expression" means the amount of CSP that can be measured by any method known in the art or the level of translation of a CSG CSNA that can be measured by any method known in the art.

The present invention provides methods for diagnosing colon cancer in a patient, in particular squamous cell carcinoma, by analyzing for changes in levels of CSNA or CSP in cells, tissues, organs or bodily fluids compared with levels of CSNA or CSP in cells, tissues, organs or bodily fluids of preferably the same type from a normal human control, wherein an increase, or decrease in certain cases, in levels of a CSNA or CSP in the patient versus the normal human control is associated with the presence of colon cancer or with a predilection to the disease. In another preferred embodiment, the present invention provides methods for diagnosing colon cancer in a patient by analyzing changes in the structure of the mRNA of a CSG compared to the mRNA from a normal control. These changes include, without limitation, aberrant splicing, alterations in polyadenylation and/or alterations in 5' nucleotide capping. In yet another preferred embodiment, the present invention provides methods for diagnosing colon cancer in a patient by analyzing changes in a CSP compared to a CSP from a normal control. These changes include, *e.g.*, alterations in glycosylation and/or phosphorylation of the CSP or subcellular CSP localization.

In a preferred embodiment, the expression of a CSNA is measured by determining the amount of an mRNA that encodes an amino acid sequence selected from

SEQ ID NO: 101 through 176, a homolog, an allelic variant, or a fragment thereof. In a more preferred embodiment, the CSNA expression that is measured is the level of expression of a CSNA mRNA selected from SEQ ID NO: 1 through 100, or a hybridizing nucleic acid, homologous nucleic acid or allelic variant thereof, or a part of any of these nucleic acids. CSNA expression may be measured by any method known in the art, such as those described *supra*, including measuring mRNA expression by Northern blot, quantitative or qualitative reverse transcriptase PCR (RT-PCR), microarray, dot or slot blots or *in situ* hybridization. *See, e.g.*, Ausubel (1992), *supra*; Ausubel (1999), *supra*; Sambrook (1989), *supra*; and Sambrook (2001), *supra*. CSNA transcription may be measured by any method known in the art including using a reporter gene hooked up to the promoter of a CSG of interest or doing nuclear run-off assays. Alterations in mRNA structure, *e.g.*, aberrant splicing variants, may be determined by any method known in the art, including, RT-PCR followed by sequencing or restriction analysis. As necessary, CSNA expression may be compared to a known control, such as normal colon nucleic acid, to detect a change in expression.

In another preferred embodiment, the expression of a CSP is measured by determining the level of a CSP having an amino acid sequence selected from the group consisting of SEQ ID NO: 101 through 176, a homolog, an allelic variant, or a fragment thereof. Such levels are preferably determined in at least one of cells, tissues, organs and/or bodily fluids, including determination of normal and abnormal levels. Thus, for instance, a diagnostic assay in accordance with the invention for diagnosing over- or underexpression of CSNA or CSP compared to normal control bodily fluids, cells, or tissue samples may be used to diagnose the presence of colon cancer. The expression level of a CSP may be determined by any method known in the art, such as those described *supra*. In a preferred embodiment, the CSP expression level may be determined by radioimmunoassays, competitive-binding assays, ELISA, Western blot, FACS, immunohistochemistry, immunoprecipitation, proteomic approaches: two-dimensional gel electrophoresis (2D electrophoresis) and non-gel-based approaches such as mass spectrometry or protein interaction profiling. *See*, e.g, Harlow (1999), *supra*; Ausubel (1992), *supra*; and Ausubel (1999), *supra*. Alterations in the CSP structure may be determined by any method known in the art, including, *e.g.*, using antibodies that specifically recognize phosphoserine, phosphothreonine or

phosphotyrosine residues, two-dimensional polyacrylamide gel electrophoresis (2D PAGE) and/or chemical analysis of amino acid residues of the protein. *Id.*

In a preferred embodiment, a radioimmunoassay (RIA) or an ELISA is used. An antibody specific to a CSP is prepared if one is not already available. In a preferred embodiment, the antibody is a monoclonal antibody. The anti-CSP antibody is bound to a solid support and any free protein binding sites on the solid support are blocked with a protein such as bovine serum albumin. A sample of interest is incubated with the antibody on the solid support under conditions in which the CSP will bind to the anti-CSP antibody. The sample is removed, the solid support is washed to remove unbound material, and an anti-CSP antibody that is linked to a detectable reagent (a radioactive substance for RIA and an enzyme for ELISA) is added to the solid support and incubated under conditions in which binding of the CSP to the labeled antibody will occur. After binding, the unbound labeled antibody is removed by washing. For an ELISA, one or more substrates are added to produce a colored reaction product that is based upon the amount of a CSP in the sample. For an RIA, the solid support is counted for radioactive decay signals by any method known in the art. Quantitative results for both RIA and ELISA typically are obtained by reference to a standard curve.

Other methods to measure CSP levels are known in the art. For instance, a competition assay may be employed wherein an anti-CSP antibody is attached to a solid support and an allocated amount of a labeled CSP and a sample of interest are incubated with the solid support. The amount of labeled CSP detected which is attached to the solid support can be correlated to the quantity of a CSP in the sample.

Of the proteomic approaches, 2D PAGE is a well-known technique. Isolation of individual proteins from a sample such as serum is accomplished using sequential separation of proteins by isoelectric point and molecular weight. Typically, polypeptides are first separated by isoelectric point (the first dimension) and then separated by size using an electric current (the second dimension). In general, the second dimension is perpendicular to the first dimension. Because no two proteins with different sequences are identical on the basis of both size and charge, the result of 2D PAGE is a roughly square gel in which each protein occupies a unique spot. Analysis of the spots with chemical or antibody probes, or subsequent protein microsequencing can reveal the relative abundance of a given protein and the identity of the proteins in the sample.

Expression levels of a CSNA can be determined by any method known in the art, including PCR and other nucleic acid methods, such as ligase chain reaction (LCR) and nucleic acid sequence based amplification (NASBA), can be used to detect malignant cells for diagnosis and monitoring of various malignancies. For example, reverse-transcriptase PCR (RT-PCR) is a powerful technique which can be used to detect the presence of a specific mRNA population in a complex mixture of thousands of other mRNA species. In RT-PCR, an mRNA species is first reverse transcribed to complementary DNA (cDNA) with use of the enzyme reverse transcriptase; the cDNA is then amplified as in a standard PCR reaction.

Hybridization to specific DNA molecules (*e.g.*, oligonucleotides) arrayed on a solid support can be used to both detect the expression of and quantitate the level of expression of one or more CSNAs of interest. In this approach, all or a portion of one or more CSNAs is fixed to a substrate. A sample of interest, which may comprise RNA, *e.g.*, total RNA or polyA-selected mRNA, or a complementary DNA (cDNA) copy of the RNA is incubated with the solid support under conditions in which hybridization will occur between the DNA on the solid support and the nucleic acid molecules in the sample of interest. Hybridization between the substrate-bound DNA and the nucleic acid molecules in the sample can be detected and quantitated by several means, including, without limitation, radioactive labeling or fluorescent labeling of the nucleic acid molecule or a secondary molecule designed to detect the hybrid.

The above tests can be carried out on samples derived from a variety of cells, bodily fluids and/or tissue extracts such as homogenates or solubilized tissue obtained from a patient. Tissue extracts are obtained routinely from tissue biopsy and autopsy material. Bodily fluids useful in the present invention include blood, urine, saliva or any other bodily secretion or derivative thereof. By blood it is meant to include whole blood, plasma, serum or any derivative of blood. In a preferred embodiment, the specimen tested for expression of CSNA or CSP includes, without limitation, colon tissue, fluid obtained by bronchial alveolar lavage (BAL), sputum, colon cells grown in cell culture, blood, serum, lymph node tissue and lymphatic fluid. In another preferred embodiment, especially when metastasis of a primary colon cancer is known or suspected, specimens include, without limitation, tissues from brain, bone, bone marrow, liver, adrenal glands and colon. In general, the tissues may be sampled by biopsy, including, without

limitation, needle biopsy, *e.g.*, transthoracic needle aspiration, cervical mediatinoscopy, endoscopic lymph node biopsy, video-assisted thoracoscopy, exploratory thoracotomy, bone marrow biopsy and bone marrow aspiration. *See* Scott, *supra* and Franklin, pp. 529-570, in Kane, *supra*. For early and inexpensive detection, assaying for changes in CSNAs or CSPs in cells in sputum samples may be particularly useful. Methods of obtaining and analyzing sputum samples is disclosed in Franklin, *supra*.

All the methods of the present invention may optionally include determining the expression levels of one or more other cancer markers in addition to determining the expression level of a CSNA or CSP. In many cases, the use of another cancer marker will decrease the likelihood of false positives or false negatives. In one embodiment, the one or more other cancer markers include other CSNA or CSPs as disclosed herein. Other cancer markers useful in the present invention will depend on the cancer being tested and are known to those of skill in the art. In a preferred embodiment, at least one other cancer marker in addition to a particular CSNA or CSP is measured. In a more preferred embodiment, at least two other additional cancer markers are used. In an even more preferred embodiment, at least three, more preferably at least five, even more preferably at least ten additional cancer markers are used.

Diagnosing

In one aspect, the invention provides a method for determining the expression levels and/or structural alterations of one or more CSNAs and/or CSPs in a sample from a patient suspected of having colon cancer. In general, the method comprises the steps of obtaining the sample from the patient, determining the expression level or structural alterations of a CSNA and/or CSP and then ascertaining whether the patient has colon cancer from the expression level of the CSNA or CSP. In general, if high expression relative to a control of a CSNA or CSP is indicative of colon cancer, a diagnostic assay is considered positive if the level of expression of the CSNA or CSP is at least two times higher, and more preferably are at least five times higher, even more preferably at least ten times higher, than in preferably the same cells, tissues or bodily fluid of a normal human control. In contrast, if low expression relative to a control of a CSNA or CSP is indicative of colon cancer, a diagnostic assay is considered positive if the level of expression of the CSNA or CSP is at least two times lower, more preferably are at least five times lower, even more preferably at least ten times lower than in preferably the

same cells, tissues or bodily fluid of a normal human control. The normal human control may be from a different patient or from uninvolved tissue of the same patient.

The present invention also provides a method of determining whether colon cancer has metastasized in a patient. One may identify whether the colon cancer has metastasized by measuring the expression levels and/or structural alterations of one or more CSNAs and/or CSPs in a variety of tissues. The presence of a CSNA or CSP in a certain tissue at levels higher than that of corresponding noncancerous tissue (*e.g.*, the same tissue from another individual) is indicative of metastasis if high level expression of a CSNA or CSP is associated with colon cancer. Similarly, the presence of a CSNA or CSP in a tissue at levels lower than that of corresponding noncancerous tissue is indicative of metastasis if low level expression of a CSNA or CSP is associated with colon cancer. Further, the presence of a structurally altered CSNA or CSP that is associated with colon cancer is also indicative of metastasis.

In general, if high expression relative to a control of a CSNA or CSP is indicative of metastasis, an assay for metastasis is considered positive if the level of expression of the CSNA or CSP is at least two times higher, and more preferably are at least five times higher, even more preferably at least ten times higher, than in preferably the same cells, tissues or bodily fluid of a normal human control. In contrast, if low expression relative to a control of a CSNA or CSP is indicative of metastasis, an assay for metastasis is considered positive if the level of expression of the CSNA or CSP is at least two times lower, more preferably are at least five times lower, even more preferably at least ten times lower than in preferably the same cells, tissues or bodily fluid of a normal human control.

The CSNA or CSP of this invention may be used as element in an array or a multi-analyte test to recognize expression patterns associated with colon cancers or other colon related disorders. In addition, the sequences of either the nucleic acids or proteins may be used as elements in a computer program for pattern recognition of colon disorders.

Staging

The invention also provides a method of staging colon cancer in a human patient. The method comprises identifying a human patient having colon cancer and analyzing

cells, tissues or bodily fluids from such human patient for expression levels and/or structural alterations of one or more CSNAs or CSPs. First, one or more tumors from a variety of patients are staged according to procedures well-known in the art, and the expression level of one or more CSNAs or CSPs is determined for each stage to obtain a standard expression level for each CSNA and CSP. Then, the CSNA or CSP expression levels are determined in a biological sample from a patient whose stage of cancer is not known. The CSNA or CSP expression levels from the patient are then compared to the standard expression level. By comparing the expression level of the CSNAs and CSPs from the patient to the standard expression levels, one may determine the stage of the tumor. The same procedure may be followed using structural alterations of a CSNA or CSP to determine the stage of a colon cancer.

Monitoring

Further provided is a method of monitoring colon cancer in a human patient. One may monitor a human patient to determine whether there has been metastasis and, if there has been, when metastasis began to occur. One may also monitor a human patient to determine whether a preneoplastic lesion has become cancerous. One may also monitor a human patient to determine whether a therapy, *e.g.*, chemotherapy, radiotherapy or surgery, has decreased or eliminated the colon cancer. The method comprises identifying a human patient that one wants to monitor for colon cancer, periodically analyzing cells, tissues or bodily fluids from such human patient for expression levels of one or more CSNAs or CSPs, and comparing the CSNA or CSP levels over time to those CSNA or CSP expression levels obtained previously. Patients may also be monitored by measuring one or more structural alterations in a CSNA or CSP that are associated with colon cancer.

If increased expression of a CSNA or CSP is associated with metastasis, treatment failure, or conversion of a preneoplastic lesion to a cancerous lesion, then detecting an increase in the expression level of a CSNA or CSP indicates that the tumor is metastasizing, that treatment has failed or that the lesion is cancerous, respectively. One having ordinary skill in the art would recognize that if this were the case, then a decreased expression level would be indicative of no metastasis, effective therapy or failure to progress to a neoplastic lesion. If decreased expression of a CSNA or CSP is associated with metastasis, treatment failure, or conversion of a preneoplastic lesion to a

cancerous lesion, then detecting an decrease in the expression level of a CSNA or CSP indicates that the tumor is metastasizing, that treatment has failed or that the lesion is cancerous, respectively. In a preferred embodiment, the levels of CSNAs or CSPs are determined from the same cell type, tissue or bodily fluid as prior patient samples. Monitoring a patient for onset of colon cancer metastasis is periodic and preferably is done on a quarterly basis, but may be done more or less frequently.

The methods described herein can further be utilized as prognostic assays to identify subjects having or at risk of developing a disease or disorder associated with increased or decreased expression levels of a CSNA and/or CSP. The present invention provides a method in which a test sample is obtained from a human patient and one or more CSNAs and/or CSPs are detected. The presence of higher (or lower) CSNA or CSP levels as compared to normal human controls is diagnostic for the human patient being at risk for developing cancer, particularly colon cancer. The effectiveness of therapeutic agents to decrease (or increase) expression or activity of one or more CSNAs and/or CSPs of the invention can also be monitored by analyzing levels of expression of the CSNAs and/or CSPs in a human patient in clinical trials or in *in vitro* screening assays such as in human cells. In this way, the gene expression pattern can serve as a marker, indicative of the physiological response of the human patient or cells, as the case may be, to the agent being tested.

Detection of Genetic Lesions or Mutations

The methods of the present invention can also be used to detect genetic lesions or mutations in a CSG, thereby determining if a human with the genetic lesion is susceptible to developing colon cancer or to determine what genetic lesions are responsible, or are partly responsible, for a person's existing colon cancer. Genetic lesions can be detected, for example, by ascertaining the existence of a deletion, insertion and/or substitution of one or more nucleotides from the CSGs of this invention, a chromosomal rearrangement of CSG, an aberrant modification of CSG (such as of the methylation pattern of the genomic DNA), or allelic loss of a CSG. Methods to detect such lesions in the CSG of this invention are known to those having ordinary skill in the art following the teachings of the specification.

Methods of Detecting Noncancerous Colon Diseases

The invention also provides a method for determining the expression levels and/or structural alterations of one or more CSNAs and/or CSPs in a sample from a patient suspected of having or known to have a noncancerous colon disease. In general, the method comprises the steps of obtaining a sample from the patient, determining the expression level or structural alterations of a CSNA and/or CSP, comparing the expression level or structural alteration of the CSNA or CSP to a normal colon control, and then ascertaining whether the patient has a noncancerous colon disease. In general, if high expression relative to a control of a CSNA or CSP is indicative of a particular noncancerous colon disease, a diagnostic assay is considered positive if the level of expression of the CSNA or CSP is at least two times higher, and more preferably are at least five times higher, even more preferably at least ten times higher, than in preferably the same cells, tissues or bodily fluid of a normal human control. In contrast, if low expression relative to a control of a CSNA or CSP is indicative of a noncancerous colon disease, a diagnostic assay is considered positive if the level of expression of the CSNA or CSP is at least two times lower, more preferably are at least five times lower, even more preferably at least ten times lower than in preferably the same cells, tissues or bodily fluid of a normal human control. The normal human control may be from a different patient or from uninvolved tissue of the same patient.

One having ordinary skill in the art may determine whether a CSNA and/or CSP is associated with a particular noncancerous colon disease by obtaining colon tissue from a patient having a noncancerous colon disease of interest and determining which CSNAs and/or CSPs are expressed in the tissue at either a higher or a lower level than in normal colon tissue. In another embodiment, one may determine whether a CSNA or CSP exhibits structural alterations in a particular noncancerous colon disease state by obtaining colon tissue from a patient having a noncancerous colon disease of interest and determining the structural alterations in one or more CSNAs and/or CSPs relative to normal colon tissue.

Methods for Identifying Colon Tissue

In another aspect, the invention provides methods for identifying colon tissue. These methods are particularly useful in, *e.g.*, forensic science, colon cell differentiation and development, and in tissue engineering.

In one embodiment, the invention provides a method for determining whether a sample is colon tissue or has colon tissue-like characteristics. The method comprises the steps of providing a sample suspected of comprising colon tissue or having colon tissue-like characteristics, determining whether the sample expresses one or more CSNAs and/or CSPs, and, if the sample expresses one or more CSNAs and/or CSPs, concluding that the sample comprises colon tissue. In a preferred embodiment, the CSNA encodes a polypeptide having an amino acid sequence selected from SEQ ID NO: 101 through 176, or a homolog, allelic variant or fragment thereof. In a more preferred embodiment, the CSNA has a nucleotide sequence selected from SEQ ID NO: 1 through 100, or a hybridizing nucleic acid, an allelic variant or a part thereof. Determining whether a sample expresses a CSNA can be accomplished by any method known in the art. Preferred methods include hybridization to microarrays, Northern blot hybridization, and quantitative or qualitative RT-PCR. In another preferred embodiment, the method can be practiced by determining whether a CSP is expressed. Determining whether a sample expresses a CSP can be accomplished by any method known in the art. Preferred methods include Western blot, ELISA, RIA and 2D PAGE. In one embodiment, the CSP has an amino acid sequence selected from SEQ ID NO: 101 through 176, or a homolog, allelic variant or fragment thereof. In another preferred embodiment, the expression of at least two CSNAs and/or CSPs is determined. In a more preferred embodiment, the expression of at least three, more preferably four and even more preferably five CSNAs and/or CSPs are determined.

In one embodiment, the method can be used to determine whether an unknown tissue is colon tissue. This is particularly useful in forensic science, in which small, damaged pieces of tissues that are not identifiable by microscopic or other means are recovered from a crime or accident scene. In another embodiment, the method can be used to determine whether a tissue is differentiating or developing into colon tissue. This is important in monitoring the effects of the addition of various agents to cell or tissue culture, *e.g.*, in producing new colon tissue by tissue engineering. These agents include,

e.g., growth and differentiation factors, extracellular matrix proteins and culture medium. Other factors that may be measured for effects on tissue development and differentiation include gene transfer into the cells or tissues, alterations in pH, aqueous:air interface and various other culture conditions.

Methods for Producing and Modifying Colon Tissue

In another aspect, the invention provides methods for producing engineered colon tissue or cells. In one embodiment, the method comprises the steps of providing cells, introducing a CSNA or a CSG into the cells, and growing the cells under conditions in which they exhibit one or more properties of colon tissue cells. In a preferred embodiment, the cells are pluripotent. As is well-known in the art, normal colon tissue comprises a large number of different cell types. Thus, in one embodiment, the engineered colon tissue or cells comprises one of these cell types. In another embodiment, the engineered colon tissue or cells comprises more than one colon cell type. Further, the culture conditions of the cells or tissue may require manipulation in order to achieve full differentiation and development of the colon cell tissue. Methods for manipulating culture conditions are well-known in the art.

Nucleic acid molecules encoding one or more CSPs are introduced into cells, preferably pluripotent cells. In a preferred embodiment, the nucleic acid molecules encode CSPs having amino acid sequences selected from SEQ ID NO: 101 through 176, or homologous proteins, analogs, allelic variants or fragments thereof. In a more preferred embodiment, the nucleic acid molecules have a nucleotide sequence selected from SEQ ID NO: 1 through 100, or hybridizing nucleic acids, allelic variants or parts thereof. In another highly preferred embodiment, a CSG is introduced into the cells. Expression vectors and methods of introducing nucleic acid molecules into cells are well-known in the art and are described in detail, *supra*.

Artificial colon tissue may be used to treat patients who have lost some or all of their colon function.

Pharmaceutical Compositions

In another aspect, the invention provides pharmaceutical compositions comprising the nucleic acid molecules, polypeptides, antibodies, antibody derivatives,

antibody fragments, agonists, antagonists, and inhibitors of the present invention. In a preferred embodiment, the pharmaceutical composition comprises a CSNA or part thereof. In a more preferred embodiment, the CSNA has a nucleotide sequence selected from the group consisting of SEQ ID NO: 1 through 100, a nucleic acid that hybridizes thereto, an allelic variant thereof, or a nucleic acid that has substantial sequence identity thereto. In another preferred embodiment, the pharmaceutical composition comprises a CSP or fragment thereof. In a more preferred embodiment, the CSP having an amino acid sequence that is selected from the group consisting of SEQ ID NO: 101 through 176, a polypeptide that is homologous thereto, a fusion protein comprising all or a portion of the polypeptide, or an analog or derivative thereof. In another preferred embodiment, the pharmaceutical composition comprises an anti-CSP antibody, preferably an antibody that specifically binds to a CSP having an amino acid that is selected from the group consisting of SEQ ID NO: 101 through 176, or an antibody that binds to a polypeptide that is homologous thereto, a fusion protein comprising all or a portion of the polypeptide, or an analog or derivative thereof.

Such a composition typically contains from about 0.1 to 90% by weight of a therapeutic agent of the invention formulated in and/or with a pharmaceutically acceptable carrier or excipient.

Pharmaceutical formulation is a well-established art, and is further described in Gennaro (ed.), Remington: The Science and Practice of Pharmacy, 20th ed., Lippincott, Williams & Wilkins (2000); Ansel *et al.*, Pharmaceutical Dosage Forms and Drug Delivery Systems, 7th ed., Lippincott Williams & Wilkins (1999); and Kibbe (ed.), Handbook of Pharmaceutical Excipients American Pharmaceutical Association, 3rd ed. (2000), the disclosures of which are incorporated herein by reference in their entireties, and thus need not be described in detail herein.

Briefly, formulation of the pharmaceutical compositions of the present invention will depend upon the route chosen for administration. The pharmaceutical compositions utilized in this invention can be administered by various routes including both enteral and parenteral routes, including oral, intravenous, intramuscular, subcutaneous, inhalation, topical, sublingual, rectal, intra-arterial, intramedullary, intrathecal, intraventricular, transmucosal, transdermal, intranasal, intraperitoneal, intrapulmonary, and intrauterine.

Oral dosage forms can be formulated as tablets, pills, dragees, capsules, liquids, gels, syrups, slurries, suspensions, and the like, for ingestion by the patient.

Solid formulations of the compositions for oral administration can contain suitable carriers or excipients, such as carbohydrate or protein fillers, such as sugars, including lactose, sucrose, mannitol, or sorbitol; starch from corn, wheat, rice, potato, or other plants; cellulose, such as methyl cellulose, hydroxypropylmethyl-cellulose, sodium carboxymethylcellulose, or microcrystalline cellulose; gums including arabic and tragacanth; proteins such as gelatin and collagen; inorganics, such as kaolin, calcium carbonate, dicalcium phosphate, sodium chloride; and other agents such as acacia and alginic acid.

Agents that facilitate disintegration and/or solubilization can be added, such as the cross-linked polyvinyl pyrrolidone, agar, alginic acid, or a salt thereof, such as sodium alginate, microcrystalline cellulose, corn starch, sodium starch glycolate, and alginic acid.

Tablet binders that can be used include acacia, methylcellulose, sodium carboxymethylcellulose, polyvinylpyrrolidone (Povidone™), hydroxypropyl methylcellulose, sucrose, starch and ethylcellulose.

Lubricants that can be used include magnesium stearates, stearic acid, silicone fluid, talc, waxes, oils, and colloidal silica.

Fillers, agents that facilitate disintegration and/or solubilization, tablet binders and lubricants, including the aforementioned, can be used singly or in combination.

Solid oral dosage forms need not be uniform throughout. For example, dragee cores can be used in conjunction with suitable coatings, such as concentrated sugar solutions, which can also contain gum arabic, talc, polyvinylpyrrolidone, carbopol gel, polyethylene glycol, and/or titanium dioxide, lacquer solutions, and suitable organic solvents or solvent mixtures.

Oral dosage forms of the present invention include push-fit capsules made of gelatin, as well as soft, sealed capsules made of gelatin and a coating, such as glycerol or sorbitol. Push-fit capsules can contain active ingredients mixed with a filler or binders, such as lactose or starches, lubricants, such as talc or magnesium stearate, and, optionally, stabilizers. In soft capsules, the active compounds can be dissolved or

suspended in suitable liquids, such as fatty oils, liquid, or liquid polyethylene glycol with or without stabilizers.

Additionally, dyestuffs or pigments can be added to the tablets or dragee coatings for product identification or to characterize the quantity of active compound, *i.e.*, dosage.

Liquid formulations of the pharmaceutical compositions for oral (enteral) administration are prepared in water or other aqueous vehicles and can contain various suspending agents such as methylcellulose, alginates, tragacanth, pectin, kelgin, carrageenan, acacia, polyvinylpyrrolidone, and polyvinyl alcohol. The liquid formulations can also include solutions, emulsions, syrups and elixirs containing, together with the active compound(s), wetting agents, sweeteners, and coloring and flavoring agents.

The pharmaceutical compositions of the present invention can also be formulated for parenteral administration. Formulations for parenteral administration can be in the form of aqueous or non-aqueous isotonic sterile injection solutions or suspensions.

For intravenous injection, water soluble versions of the compounds of the present invention are formulated in, or if provided as a lyophilate, mixed with, a physiologically acceptable fluid vehicle, such as 5% dextrose ("D5"), physiologically buffered saline, 0.9% saline, Hanks' solution, or Ringer's solution. Intravenous formulations may include carriers, excipients or stabilizers including, without limitation, calcium, human serum albumin, citrate, acetate, calcium chloride, carbonate, and other salts.

Intramuscular preparations, *e.g.* a sterile formulation of a suitable soluble salt form of the compounds of the present invention, can be dissolved and administered in a pharmaceutical excipient such as Water-for-Injection, 0.9% saline, or 5% glucose solution. Alternatively, a suitable insoluble form of the compound can be prepared and administered as a suspension in an aqueous base or a pharmaceutically acceptable oil base, such as an ester of a long chain fatty acid (*e.g.*, ethyl oleate), fatty oils such as sesame oil, triglycerides, or liposomes.

Parenteral formulations of the compositions can contain various carriers such as vegetable oils, dimethylacetamide, dimethylformamide, ethyl lactate, ethyl carbonate, isopropyl myristate, ethanol, polyols (glycerol, propylene glycol, liquid polyethylene glycol, and the like).

Aqueous injection suspensions can also contain substances that increase the viscosity of the suspension, such as sodium carboxymethyl cellulose, sorbitol, or dextran. Non-lipid polycationic amino polymers can also be used for delivery. Optionally, the suspension can also contain suitable stabilizers or agents that increase the solubility of the compounds to allow for the preparation of highly concentrated solutions.

Pharmaceutical compositions of the present invention can also be formulated to permit injectable, long-term, deposition. Injectable depot forms may be made by forming microencapsulated matrices of the compound in biodegradable polymers such as polylactide-polyglycolide. Depending upon the ratio of drug to polymer and the nature of the particular polymer employed, the rate of drug release can be controlled. Examples of other biodegradable polymers include poly(orthoesters) and poly(anhydrides). Depot injectable formulations are also prepared by entrapping the drug in microemulsions that are compatible with body tissues.

The pharmaceutical compositions of the present invention can be administered topically.

For topical use the compounds of the present invention can also be prepared in suitable forms to be applied to the skin, or mucus membranes of the nose and throat, and can take the form of lotions, creams, ointments, liquid sprays or inhalants, drops, tinctures, lozenges, or throat paints. Such topical formulations further can include chemical compounds such as dimethylsulfoxide (DMSO) to facilitate surface penetration of the active ingredient. In other transdermal formulations, typically in patch-delivered formulations, the pharmaceutically active compound is formulated with one or more skin penetrants, such as 2-N-methyl-pyrrolidone (NMP) or Azone. A topical semi-solid ointment formulation typically contains a concentration of the active ingredient from about 1 to 20%, *e.g.*, 5 to 10%, in a carrier such as a pharmaceutical cream base.

For application to the eyes or ears, the compounds of the present invention can be presented in liquid or semi-liquid form formulated in hydrophobic or hydrophilic bases as ointments, creams, lotions, paints or powders.

For rectal administration the compounds of the present invention can be administered in the form of suppositories admixed with conventional carriers such as cocoa butter, wax or other glyceride.

Inhalation formulations can also readily be formulated. For inhalation, various powder and liquid formulations can be prepared. For aerosol preparations, a sterile formulation of the compound or salt form of the compound may be used in inhalers, such as metered dose inhalers, and nebulizers. Aerosolized forms may be especially useful for treating respiratory disorders.

Alternatively, the compounds of the present invention can be in powder form for reconstitution in the appropriate pharmaceutically acceptable carrier at the time of delivery.

The pharmaceutically active compound in the pharmaceutical compositions of the present invention can be provided as the salt of a variety of acids, including but not limited to hydrochloric, sulfuric, acetic, lactic, tartaric, malic, and succinic acid. Salts tend to be more soluble in aqueous or other protonic solvents than are the corresponding free base forms.

After pharmaceutical compositions have been prepared, they are packaged in an appropriate container and labeled for treatment of an indicated condition.

The active compound will be present in an amount effective to achieve the intended purpose. The determination of an effective dose is well within the capability of those skilled in the art.

A "therapeutically effective dose" refers to that amount of active ingredient, for example CSP polypeptide, fusion protein, or fragments thereof, antibodies specific for CSP, agonists, antagonists or inhibitors of CSP, which ameliorates the signs or symptoms of the disease or prevents progression thereof; as would be understood in the medical arts, cure, although desired, is not required.

The therapeutically effective dose of the pharmaceutical agents of the present invention can be estimated initially by *in vitro* tests, such as cell culture assays, followed by assay in model animals, usually mice, rats, rabbits, dogs, or pigs. The animal model can also be used to determine an initial preferred concentration range and route of administration.

For example, the ED50 (the dose therapeutically effective in 50% of the population) and LD50 (the dose lethal to 50% of the population) can be determined in one or more cell culture of animal model systems. The dose ratio of toxic to therapeutic

effects is the therapeutic index, which can be expressed as LD50/ED50. Pharmaceutical compositions that exhibit large therapeutic indices are preferred.

The data obtained from cell culture assays and animal studies are used in formulating an initial dosage range for human use, and preferably provide a range of circulating concentrations that includes the ED50 with little or no toxicity. After administration, or between successive administrations, the circulating concentration of active agent varies within this range depending upon pharmacokinetic factors well-known in the art, such as the dosage form employed, sensitivity of the patient, and the route of administration.

The exact dosage will be determined by the practitioner, in light of factors specific to the subject requiring treatment. Factors that can be taken into account by the practitioner include the severity of the disease state, general health of the subject, age, weight, gender of the subject, diet, time and frequency of administration, drug combination(s), reaction sensitivities, and tolerance/response to therapy. Long-acting pharmaceutical compositions can be administered every 3 to 4 days, every week, or once every two weeks depending on half-life and clearance rate of the particular formulation.

Normal dosage amounts may vary from 0.1 to 100,000 micrograms, up to a total dose of about 1 g, depending upon the route of administration. Where the therapeutic agent is a protein or antibody of the present invention, the therapeutic protein or antibody agent typically is administered at a daily dosage of 0.01 mg to 30 mg/kg of body weight of the patient (*e.g.*, 1 mg/kg to 5 mg/kg). The pharmaceutical formulation can be administered in multiple doses per day, if desired, to achieve the total desired daily dose.

Guidance as to particular dosages and methods of delivery is provided in the literature and generally available to practitioners in the art. Those skilled in the art will employ different formulations for nucleotides than for proteins or their inhibitors. Similarly, delivery of polynucleotides or polypeptides will be specific to particular cells, conditions, locations, etc.

Conventional methods, known to those of ordinary skill in the art of medicine, can be used to administer the pharmaceutical formulation(s) of the present invention to the patient. The pharmaceutical compositions of the present invention can be administered alone, or in combination with other therapeutic agents or interventions.

Therapeutic Methods

The present invention further provides methods of treating subjects having defects in a gene of the invention , *e.g.*, in expression, activity, distribution, localization, and/or solubility, which can manifest as a disorder of colon function. As used herein, "treating" includes all medically-acceptable types of therapeutic intervention, including palliation and prophylaxis (prevention) of disease. The term "treating" encompasses any improvement of a disease, including minor improvements. These methods are discussed below.

Gene Therapy and Vaccines

The isolated nucleic acids of the present invention can also be used to drive *in vivo* expression of the polypeptides of the present invention. *In vivo* expression can be driven from a vector, typically a viral vector, often a vector based upon a replication incompetent retrovirus, an adenovirus, or an adeno-associated virus (AAV) , for purpose of gene therapy. *In vivo* expression can also be driven from signals endogenous to the nucleic acid or from a vector, often a plasmid vector, such as pVAX1 (Invitrogen, Carlsbad, CA, USA), for purpose of "naked" nucleic acid vaccination, as further described in U.S. Patents 5,589,466; 5,679,647; 5,804,566; 5,830,877; 5,843,913; 5,880,104; 5,958,891; 5,985,847; 6,017,897; 6,110,898; and 6,204,250, the disclosures of which are incorporated herein by reference in their entireties. For cancer therapy, it is preferred that the vector also be tumor-selective. *See, e.g.*, Doronin *et al.*, *J. Virol.* 75: 3314-24 (2001).

In another embodiment of the therapeutic methods of the present invention, a therapeutically effective amount of a pharmaceutical composition comprising a nucleic acid of the present invention is administered. The nucleic acid can be delivered in a vector that drives expression of a CSP, fusion protein, or fragment thereof, or without such vector. Nucleic acid compositions that can drive expression of a CSP are administered, for example, to complement a deficiency in the native CSP, or as DNA vaccines. Expression vectors derived from virus, replication deficient retroviruses, adenovirus, adeno-associated (AAV) virus, herpes virus, or vaccinia virus can be used as can plasmids. *See*, *e.g.*, Cid-Arregui, *supra*. In a preferred embodiment, the nucleic acid

molecule encodes a CSP having the amino acid sequence of SEQ ID NO: 101 through 176, or a fragment, fusion protein, allelic variant or homolog thereof.

In still other therapeutic methods of the present invention, pharmaceutical compositions comprising host cells that express a CSP, fusions, or fragments thereof can be administered. In such cases, the cells are typically autologous, so as to circumvent xenogeneic or allotypic rejection, and are administered to complement defects in CSP production or activity. In a preferred embodiment, the nucleic acid molecules in the cells encode a CSP having the amino acid sequence of SEQ ID NO: 101 through 176, or a fragment, fusion protein, allelic variant or homolog thereof.

Antisense Administration

Antisense nucleic acid compositions, or vectors that drive expression of a CSG antisense nucleic acid, are administered to downregulate transcription and/or translation of a CSG in circumstances in which excessive production, or production of aberrant protein, is the pathophysiologic basis of disease.

Antisense compositions useful in therapy can have a sequence that is complementary to coding or to noncoding regions of a CSG. For example, oligonucleotides derived from the transcription initiation site, *e.g.*, between positions -10 and +10 from the start site, are preferred.

Catalytic antisense compositions, such as ribozymes, that are capable of sequence-specific hybridization to CSG transcripts, are also useful in therapy. *See, e.g.*, Phylactou, *Adv. Drug Deliv. Rev.* 44(2-3): 97-108 (2000); Phylactou *et al.*, *Hum. Mol. Genet.* 7(10): 1649-53 (1998); Rossi, *Ciba Found. Symp.* 209: 195-204 (1997); and Sigurdsson *et al.*, *Trends Biotechnol.* 13(8): 286-9 (1995), the disclosures of which are incorporated herein by reference in their entireties.

Other nucleic acids useful in the therapeutic methods of the present invention are those that are capable of triplex helix formation in or near the CSG genomic locus. Such triplexing oligonucleotides are able to inhibit transcription. *See, e.g.*, Intody *et al.*, *Nucleic Acids Res.* 28(21): 4283-90 (2000); McGuffie *et al.*, *Cancer Res.* 60(14): 3790-9 (2000), the disclosures of which are incorporated herein by reference. Pharmaceutical compositions comprising such triplex forming oligos (TFOs) are administered in circumstances in which excessive production, or production of aberrant protein, is a pathophysiologic basis of disease.

In a preferred embodiment, the antisense molecule is derived from a nucleic acid molecule encoding a CSP, preferably a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176, or a fragment, allelic variant or homolog thereof. In a more preferred embodiment, the antisense molecule is derived from a nucleic acid molecule having a nucleotide sequence of SEQ ID NO: 1 through 100, or a part, allelic variant, substantially similar or hybridizing nucleic acid thereof.

Polypeptide Administration

In one embodiment of the therapeutic methods of the present invention, a therapeutically effective amount of a pharmaceutical composition comprising a CSP, a fusion protein, fragment, analog or derivative thereof is administered to a subject with a clinically-significant CSP defect.

Protein compositions are administered, for example, to complement a deficiency in native CSP. In other embodiments, protein compositions are administered as a vaccine to elicit a humoral and/or cellular immune response to CSP. The immune response can be used to modulate activity of CSP or, depending on the immunogen, to immunize against aberrant or aberrantly expressed forms, such as mutant or inappropriately expressed isoforms. In yet other embodiments, protein fusions having a toxic moiety are administered to ablate cells that aberrantly accumulate CSP.

In a preferred embodiment, the polypeptide is a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176, or a fusion protein, allelic variant, homolog, analog or derivative thereof. In a more preferred embodiment, the polypeptide is encoded by a nucleic acid molecule having a nucleotide sequence of SEQ ID NO: 1 through 100, or a part, allelic variant, substantially similar or hybridizing nucleic acid thereof.

Antibody, Agonist and Antagonist Administration

In another embodiment of the therapeutic methods of the present invention, a therapeutically effective amount of a pharmaceutical composition comprising an antibody (including fragment or derivative thereof) of the present invention is administered. As is well-known, antibody compositions are administered, for example, to antagonize activity of CSP, or to target therapeutic agents to sites of CSP presence and/or accumulation. In a preferred embodiment, the antibody specifically binds to a

CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176, or a fusion protein, allelic variant, homolog, analog or derivative thereof. In a more preferred embodiment, the antibody specifically binds to a CSP encoded by a nucleic acid molecule having a nucleotide sequence of SEQ ID NO: 1 through 100, or a part, allelic variant, substantially similar or hybridizing nucleic acid thereof.

The present invention also provides methods for identifying modulators which bind to a CSP or have a modulatory effect on the expression or activity of a CSP. Modulators which decrease the expression or activity of CSP (antagonists) are believed to be useful in treating colon cancer. Such screening assays are known to those of skill in the art and include, without limitation, cell-based assays and cell-free assays. Small molecules predicted via computer imaging to specifically bind to regions of a CSP can also be designed, synthesized and tested for use in the imaging and treatment of colon cancer. Further, libraries of molecules can be screened for potential anticancer agents by assessing the ability of the molecule to bind to the CSPs identified herein. Molecules identified in the library as being capable of binding to a CSP are key candidates for further evaluation for use in the treatment of colon cancer. In a preferred embodiment, these molecules will downregulate expression and/or activity of a CSP in cells.

In another embodiment of the therapeutic methods of the present invention, a pharmaceutical composition comprising a non-antibody antagonist of CSP is administered. Antagonists of CSP can be produced using methods generally known in the art. In particular, purified CSP can be used to screen libraries of pharmaceutical agents, often combinatorial libraries of small molecules, to identify those that specifically bind and antagonize at least one activity of a CSP.

In other embodiments a pharmaceutical composition comprising an agonist of a CSP is administered. Agonists can be identified using methods analogous to those used to identify antagonists.

In a preferred embodiment, the antagonist or agonist specifically binds to and antagonizes or agonizes, respectively, a CSP comprising an amino acid sequence of SEQ ID NO: 101 through 176, or a fusion protein, allelic variant, homolog, analog or derivative thereof. In a more preferred embodiment, the antagonist or agonist specifically binds to and antagonizes or agonizes, respectively, a CSP encoded by a

nucleic acid molecule having a nucleotide sequence of SEQ ID NO: 1 through 100, or a part, allelic variant, substantially similar or hybridizing nucleic acid thereof.

Targeting Colon Tissue

The invention also provides a method in which a polypeptide of the invention, or an antibody thereto, is linked to a therapeutic agent such that it can be delivered to the colon or to specific cells in the colon. In a preferred embodiment, an anti-CSP antibody is linked to a therapeutic agent and is administered to a patient in need of such therapeutic agent. The therapeutic agent may be a toxin, if colon tissue needs to be selectively destroyed. This would be useful for targeting and killing colon cancer cells. In another embodiment, the therapeutic agent may be a growth or differentiation factor, which would be useful for promoting colon cell function.

In another embodiment, an anti-CSP antibody may be linked to an imaging agent that can be detected using, *e.g.*, magnetic resonance imaging, CT or PET. This would be useful for determining and monitoring colon function, identifying colon cancer tumors, and identifying noncancerous colon diseases.

EXAMPLES

Example 1: Gene Expression analysis

CSGs were identified by a systematic analysis of gene expression data in the LIFESEQ® Gold database available from Incyte Genomics Inc (Palo Alto, CA) using the data mining software package CLASP™ (Candidate Lead Automatic Search Program). CLASP™ is a set of algorithms that interrogate Incyte's database to identify genes that are both specific to particular tissue types as well as differentially expressed in tissues from patients with cancer. LifeSeq® Gold contains information about which genes are expressed in various tissues in the body and about the dynamics of expression in both normal and diseased states. CLASP™ first sorts the LifeSeq® Gold database into defined tissue types, such as colon, ovary and prostate. CLASP™ categorizes each tissue sample by disease state. Disease states include "healthy," "cancer," "associated with cancer," "other disease" and "other." Categorizing the disease states improves our ability to identify tissue and cancer-specific molecular targets. CLASP™ then performs a simultaneous parallel search for genes that are expressed both (1) selectively in the defined tissue type compared to other tissue types and (2) differentially in the "cancer"

disease state compared to the other disease states affecting the same, or different, tissues. This sorting is accomplished by using mathematical and statistical filters that specify the minimum change in expression levels and the minimum frequency that the differential expression pattern must be observed across the tissue samples for the gene to be considered statistically significant. The CLASP™ algorithm quantifies the relative abundance of a particular gene in each tissue type and in each disease state.

To find the CSGs of this invention, the following specific CLASP™ profiles were utilized: tissue-specific expression (CLASP 1), detectable expression only in cancer tissue (CLASP 2), highest differential expression for a given cancer (CLASP 4); differential expression in cancer tissue (CLASP 5), and. cDNA libraries were divided into 60 unique tissue types (early versions of LifeSeq® had 48 tissue types). Genes or ESTs were grouped into "gene bins," where each bin is a cluster of sequences grouped together where they share a common contig. The expression level for each gene bin was calculated for each tissue type. Differential expression significance was calculated with rigorous statistical significant testing taking into account variations in sample size and relative gene abundance in different libraries and within each library (for the equations used to determine statistically significant expression see Audic and Claverie "The significance of digital gene expression profiles," Genome Res 7(10): 986-995 (1997), including Equation 1 on page 987 and Equation 2 on page 988, the contents of which are incorporated by reference). Differentially expressed tissue-specific genes were selected based on the percentage abundance level in the targeted tissue versus all the other tissues (tissue-specificity). The expression levels for each gene in libraries of normal tissues or non-tumor tissues from cancer patients were compared with the expression levels in tissue libraries associated with tumor or disease (cancer-specificity). The results were analyzed for statistical significance.

The selection of the target genes meeting the rigorous CLASP™ profile criteria were as follows:

(a) CLASP 1: tissue-specific expression: To qualify as a CLASP 1 candidate, a gene must exhibit statistically significant expression in the tissue of interest compared to all other tissues. Only if the gene exhibits such differential expression with a 90% of confidence level is it selected as a CLASP 1 candidate.

(b) CLASP 2: detectable expression only in cancer tissue: To qualify as a CLASP 2 candidate, a gene must exhibit detectable expression in tumor tissues and undetectable expression in libraries from normal individuals and libraries from normal tissue obtained from diseased patients. In addition, such a gene must also exhibit further specificity for the tumor tissues of interest.

(c) CLASP 5: differential expression in cancer tissue: To qualify as a CLASP 5 candidate, a gene must be differentially expressed in tumor libraries in the tissue of interest compared to normal libraries for all tissues. Only if the gene exhibits such differential expression with a 90% of confidence level is it selected as a CLASP 5 candidate.

The CLASP™ scores for SEQ ID NO: 1-100 are listed below:

SEQ ID NO: 1	DEX0255_1	CLASP2
SEQ ID NO: 2	DEX0255_2	CLASP2
SEQ ID NO: 3	DEX0255_3	CLASP2
SEQ ID NO: 4	DEX0255_4	CLASP2 CLASP1
SEQ ID NO: 5	DEX0255_5	CLASP2
SEQ ID NO: 6	DEX0255_6	CLASP2
SEQ ID NO: 7	DEX0255_7	CLASP2
SEQ ID NO: 8	DEX0255_8	CLASP2
SEQ ID NO: 9	DEX0255_9	CLASP2
SEQ ID NO: 10	DEX0255_10	CLASP2
SEQ ID NO: 11	DEX0255_11	CLASP2
SEQ ID NO: 12	DEX0255_12	CLASP2
SEQ ID NO: 13	DEX0255_13	CLASP2
SEQ ID NO: 14	DEX0255_14	CLASP5 CLASP1
SEQ ID NO: 15	DEX0255_15	CLASP5 CLASP1
SEQ ID NO: 16	DEX0255_16	CLASP2
SEQ ID NO: 17	DEX0255_17	CLASP2
SEQ ID NO: 18	DEX0255_18	CLASP5 CLASP1
SEQ ID NO: 19	DEX0255_19	CLASP5 CLASP1
SEQ ID NO: 20	DEX0255_20	CLASP5 CLASP1
SEQ ID NO: 21	DEX0255_21	CLASP5 CLASP1
SEQ ID NO: 22	DEX0255_22	CLASP2
SEQ ID NO: 23	DEX0255_23	CLASP2
SEQ ID NO: 24	DEX0255_24	CLASP2 CLASP1
SEQ ID NO: 25	DEX0255_25	CLASP2 CLASP1
SEQ ID NO: 26	DEX0255_26	CLASP2
SEQ ID NO: 27	DEX0255_27	CLASP2
SEQ ID NO: 28	DEX0255_28	CLASP2 CLASP1
SEQ ID NO: 29	DEX0255_29	CLASP2 CLASP1
SEQ ID NO: 30	DEX0255_30	CLASP5 CLASP1

SEQ ID NO: 31	DEX0255_31	CLASP2
SEQ ID NO: 32	DEX0255_32	CLASP5 CLASP1
SEQ ID NO: 33	DEX0255_33	CLASP5 CLASP1
SEQ ID NO: 34	DEX0255_34	CLASP2
SEQ ID NO: 35	DEX0255_35	CLASP2
SEQ ID NO: 36	DEX0255_36	CLASP2
SEQ ID NO: 37	DEX0255_37	CLASP2
SEQ ID NO: 39	DEX0255_39	CLASP2
SEQ ID NO: 40	DEX0255_40	CLASP2
SEQ ID NO: 41	DEX0255_41	CLASP2
SEQ ID NO: 42	DEX0255_42	CLASP2
SEQ ID NO: 43	DEX0255_43	CLASP2
SEQ ID NO: 44	DEX0255_44	CLASP2
SEQ ID NO: 45	DEX0255_45	CLASP2
SEQ ID NO: 46	DEX0255_46	CLASP2 CLASP1
SEQ ID NO: 47	DEX0255_47	CLASP2
SEQ ID NO: 48	DEX0255_48	CLASP2
SEQ ID NO: 49	DEX0255_49	CLASP2
SEQ ID NO: 50	DEX0255_50	CLASP2
SEQ ID NO: 51	DEX0255_51	CLASP2
SEQ ID NO: 52	DEX0255_52	CLASP2
SEQ ID NO: 53	DEX0255_53	CLASP2
SEQ ID NO: 54	DEX0255_54	CLASP2
SEQ ID NO: 55	DEX0255_55	CLASP2
SEQ ID NO: 56	DEX0255_56	CLASP2
SEQ ID NO: 57	DEX0255_57	CLASP2 CLASP1
SEQ ID NO: 58	DEX0255_58	CLASP2 CLASP1
SEQ ID NO: 59	DEX0255_59	CLASP2
SEQ ID NO: 60	DEX0255_60	CLASP2
SEQ ID NO: 61	DEX0255_61	CLASP2
SEQ ID NO: 62	DEX0255_62	CLASP2
SEQ ID NO: 63	DEX0255_63	CLASP2
SEQ ID NO: 64	DEX0255_64	CLASP5 CLASP1
SEQ ID NO: 65	DEX0255_65	CLASP5 CLASP1
SEQ ID NO: 66	DEX0255_66	CLASP5 CLASP1
SEQ ID NO: 67	DEX0255_67	CLASP5 CLASP1
SEQ ID NO: 68	DEX0255_68	CLASP2
SEQ ID NO: 69	DEX0255_69	CLASP2
SEQ ID NO: 70	DEX0255_70	CLASP2
SEQ ID NO: 71	DEX0255_71	CLASP5 CLASP1
SEQ ID NO: 72	DEX0255_72	CLASP2
SEQ ID NO: 73	DEX0255_73	CLASP2
SEQ ID NO: 74	DEX0255_74	CLASP2
SEQ ID NO: 75	DEX0255_75	CLASP2
SEQ ID NO: 76	DEX0255_76	CLASP2
SEQ ID NO: 77	DEX0255_77	CLASP2
SEQ ID NO: 78	DEX0255_78	CLASP2
SEQ ID NO: 79	DEX0255_79	CLASP2

SEQ ID NO: 80	DEX0255_80	CLASP2
SEQ ID NO: 81	DEX0255_81	CLASP2
SEQ ID NO: 83	DEX0255_83	CLASP2
SEQ ID NO: 84	DEX0255_84	CLASP2
SEQ ID NO: 85	DEX0255_85	CLASP2
SEQ ID NO: 87	DEX0255_87	CLASP2 CLASP1
SEQ ID NO: 88	DEX0255_88	CLASP2 CLASP1
SEQ ID NO: 91	DEX0255_91	CLASP2
SEQ ID NO: 92	DEX0255_92	CLASP2
SEQ ID NO: 93	DEX0255_93	CLASP2
SEQ ID NO: 94	DEX0255_94	CLASP2
SEQ ID NO: 95	DEX0255_95	CLASP2 CLASP1
SEQ ID NO: 96	DEX0255_96	CLASP5 CLASP1
SEQ ID NO: 97	DEX0255_97	CLASP5 CLASP1
SEQ ID NO: 98	DEX0255_98	CLASP5 CLASP1
SEQ ID NO: 99	DEX0255_99	CLASP5 CLASP1
SEQ ID NO: 100	DEX0255_100	CLASP5 CLASP1

Example 2: Relative Quantitation of Gene Expression

Real-Time quantitative PCR with fluorescent Taqman probes is a quantitation detection system utilizing the 5'- 3' nuclease activity of Taq DNA polymerase. The method uses an internal fluorescent oligonucleotide probe (Taqman) labeled with a 5' reporter dye and a downstream, 3' quencher dye. During PCR, the 5'-3' nuclease activity of Taq DNA polymerase releases the reporter, whose fluorescence can then be detected by the laser detector of the Model 7700 Sequence Detection System (PE Applied Biosystems, Foster City, CA, USA). Amplification of an endogenous control is used to standardize the amount of sample RNA added to the reaction and normalize for Reverse Transcriptase (RT) efficiency. Either cyclophilin, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), ATPase, or 18S ribosomal RNA (rRNA) is used as this endogenous control. To calculate relative quantitation between all the samples studied, the target RNA levels for one sample were used as the basis for comparative results (calibrator). Quantitation relative to the "calibrator" can be obtained using the standard curve method or the comparative method (User Bulletin #2: ABI PRISM 7700 Sequence Detection System).

The tissue distribution and the level of the target gene are evaluated for every sample in normal and cancer tissues. Total RNA is extracted from normal tissues, cancer tissues, and from cancers and the corresponding matched adjacent tissues. Subsequently, first strand cDNA is prepared with reverse transcriptase and the polymerase chain

reaction is done using primers and Taqman probes specific to each target gene. The results are analyzed using the ABI PRISM 7700 Sequence Detector. The absolute numbers are relative levels of expression of the target gene in a particular tissue compared to the calibrator tissue.

One of ordinary skill can design appropriate primers. The relative levels of expression of the CSNA versus normal tissues and other cancer tissues can then be determined. All the values are compared to normal thymus (calibrator). These RNA samples are commercially available pools, originated by pooling samples of a particular tissue from different individuals.

The relative levels of expression of the CSNA in pairs of matching samples and 1 cancer and 1 normal/normal adjacent of tissue may also be determined. All the values are compared to normal thymus (calibrator). A matching pair is formed by mRNA from the cancer sample for a particular tissue and mRNA from the normal adjacent sample for that same tissue from the same individual.

In the analysis of matching samples, the CSNAs that show a high degree of tissue specificity for the tissue of interest. These results confirm the tissue specificity results obtained with normal pooled samples.

Further, the level of mRNA expression in cancer samples and the isogenic normal adjacent tissue from the same individual are compared. This comparison provides an indication of specificity for the cancer stage (*e.g.* higher levels of mRNA expression in the cancer sample compared to the normal adjacent).

Altogether, the high level of tissue specificity, plus the mRNA overexpression in matching samples tested are indicative of SEQ ID NO: 1 through 100 being a diagnostic marker for cancer.

Example 3: Protein Expression

The CSNA is amplified by polymerase chain reaction (PCR) and the amplified DNA fragment encoding the CSNA is subcloned in pET-21d for expression in *E. coli*. In addition to the CSNA coding sequence, codons for two amino acids, Met-Ala, flanking the NH_2-terminus of the coding sequence of CSNA, and six histidines, flanking the COOH-terminus of the coding sequence of CSNA, are incorporated to serve as initiating Met/restriction site and purification tag, respectively.

An over-expressed protein band of the appropriate molecular weight may be observed on a Coomassie blue stained polyacrylamide gel. This protein band is confirmed by Western blot analysis using monoclonal antibody against 6X Histidine tag.

Large-scale purification of CSP was achieved using cell paste generated from 6-liter bacterial cultures, and purified using immobilized metal affinity chromatography (IMAC). Soluble fractions that had been separated from total cell lysate were incubated with a nickle chelating resin. The column was packed and washed with five column volumes of wash buffer. CSP was eluted stepwise with various concentration imidazole buffers.

Example 4: Protein Fusions

Briefly, the human Fc portion of the IgG molecule can be PCR amplified, using primers that span the 5'and 3' ends of the sequence described below. These primers also should have convenient restriction enzyme sites that will facilitate cloning into an expression vector, preferably a mammalian expression vector. For example, if pC4 (Accession No. 209646) is used, the human Fc portion can be ligated into the BamHI cloning site. Note that the 3' BamHI site should be destroyed. Next, the vector containing the human Fc portion is re-restricted with BamHI, linearizing the vector, and a polynucleotide of the present invention, isolated by the PCR protocol described in Example 2, is ligated into this BamHI site. Note that the polynucleotide is cloned without a stop codon, otherwise a fusion protein will not be produced. If the naturally occurring signal sequence is used to produce the secreted protein, pC4 does not need a second signal peptide. Alternatively, if the naturally occurring signal sequence is not used, the vector can be modified to include a heterologous signal sequence. *See, e. g.*, WO 96/34891.

Example 5: Production of an Antibody from a Polypeptide

In general, such procedures involve immunizing an animal (preferably a mouse) with polypeptide or, more preferably, with a secreted polypeptide-expressing cell. Such cells may be cultured in any suitable tissue culture medium; however, it is preferable to culture cells in Earle's modified Eagle's medium supplemented with 10% fetal bovine serum (inactivated at about 56°C), and supplemented with about 10 g/l of nonessential amino acids, about 1,000 U/ml of penicillin, and about 100, μg/ml of streptomycin. The

splenocytes of such mice are extracted and fused with a suitable myeloma cell line. Any suitable myeloma cell line may be employed in accordance with the present invention; however, it is preferable to employ the parent myeloma cell line (SP20), available from the ATCC. After fusion, the resulting hybridoma cells are selectively maintained in HAT medium, and then cloned by limiting dilution as described by Wands *et al.*, *Gastroenterology* 80: 225-232 (1981).

The hybridoma cells obtained through such a selection are then assayed to identify clones which secrete antibodies capable of binding the polypeptide. Alternatively, additional antibodies capable of binding to the polypeptide can be produced in a two-step procedure using anti-idiotypic antibodies. Such a method makes use of the fact that antibodies are themselves antigens, and therefore, it is possible to obtain an antibody which binds to a second antibody. In accordance with this method, protein specific antibodies are used to immunize an animal, preferably a mouse. The splenocytes of such an animal are then used to produce hybridoma cells, and the hybridoma cells are screened to identify clones which produce an antibody whose ability to bind to the protein-specific antibody can be blocked by the polypeptide. Such antibodies comprise anti-idiotypic antibodies to the protein specific antibody and can be used to immunize an animal to induce formation of further protein-specific antibodies. Using the Jameson-Wolf methods the following epitopes were predicted. (Jameson and Wolf, CABIOS, 4(1), 181-186, 1988, the contents of which are incorporated by reference).

DEX0255_105	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
38-48	1.01	11
DEX0255_108	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
35-46	1.06	12
3-24	1.02	22
DEX0255_110	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
7-23	1.17	17
DEX0255_111	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
32-43	0.95	12
DEX0255_115	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
114-124	1.04	11
61-112	1.03	52

DEX0255_116	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
27-41	1.10	15
DEX0255_120	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
12-78	1.02	67
DEX0255_123	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
46-59	1.09	14
DEX0255_126	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
3-23	1.15	21
DEX0255_127	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
35-49	1.10	15
DEX0255_136	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
4-31	0.98	28
DEX0255_138	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
28-54	0.92	27
DEX0255_139	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
18-44	1.06	27
DEX0255_140	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
37-61	1.03	25
DEX0255_141	Antigenicity Index(Jameson-Wolf)	
positions	AI	avg length
240-253	1.22	14
780-793	1.22	14
813-822	1.21	10
542-551	1.21	10
273-282	1.17	10
335-344	1.12	10
509-523	1.12	15
110-161	1.10	52
204-222	1.10	19
635-721	1.09	87
366-450	1.09	85
723-740	1.08	18
452-469	1.07	18
285-303	1.07	19
834-846	1.04	13
43-90	1.02	48
92-105	0.97	14
744-762	0.97	19
314-327	0.96	14

	473-491	0.94	19
	583-596	0.93	14
	554-572	0.93	19
DEX0255_145		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	11-24	1.17	14
DEX0255_146		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	16-35	0.97	20
DEX0255_147		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	18-32	1.08	15
DEX0255_152		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	3-16	0.99	14
	38-49	0.98	12
DEX0255_154		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	22-34	1.07	13
DEX0255_157		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	24-45	1.00	22
	5-14	0.97	10
DEX0255_158		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	13-40	1.02	28
DEX0255_166		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	21-30	1.01	10
DEX0255_168		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	68-81	1.01	14
DEX0255_171		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	41-77	0.94	37
DEX0255_175		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	434-448	1.08	15
	120-129	0.92	10
	23-87	0.92	65
DEX0255_176		Antigenicity Index(Jameson-Wolf)	
	positions	AI	avg length
	5-32	1.11	28

Examples of post-translational modifications (PTMs) of the CSPs of this invention are listed below. In addition, antibodies that specifically bind such post-

translational modifications may be useful as a diagnostic or as therapeutic. Using the ProSite database (Bairoch et al., Nucleic Acids Res. 25(1):217-221 (1997), the contents of which are incorporated by reference), the following PTMs were predicted for the CSPs of the invention (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_prosite.html most recently accessed October 23, 2001). For full definitions of the PTMs see http://www.expasy.org/cgi-bin/prosite-list.pl most recently accessed October 23, 2001.

DEX0255_105	Ck2_Phospho_Site 11-14; Pkc_Phospho_Site 2-4;3-5;11-13;
DEX0255_106	Asn_Glycosylation 51-54; Ck2_Phospho_Site 35-38; Pkc_Phospho_Site 45-47;
DEX0255_107	Ck2_Phospho_Site 27-30; Pkc_Phospho_Site 13-15;
DEX0255_108	Myristyl 38-43; Pkc_Phospho_Site 14-16;28-30;
DEX0255_109	Ck2_Phospho_Site 56-59; Myristyl 15-20; Pkc_Phospho_Site 56-58;
DEX0255_110	Asn_Glycosylation 40-43; Ck2_Phospho_Site 10-13;14-17;19-22; Myristyl 15-20; Pkc_Phospho_Site 79-81;
DEX0255_111	Pkc_Phospho_Site 8-10;
DEX0255_112	Pkc_Phospho_Site 21-23;43-45;
DEX0255_114	Asn_Glycosylation 5-8; Ck2_Phospho_Site 39-42; Pkc_Phospho_Site 50-52;
DEX0255_115	Asn_Glycosylation 2-5; Glycosaminoglycan 19-22; Myristyl 33-38;37-42;38-43;40-45;49-54;55-60;58-63;64-69;84-89; Pkc_Phospho_Site 88-90;143-145; Prokar_Lipoprotein 31-41;
DEX0255_116	Myristyl 63-68;
DEX0255_117	Asn_Glycosylation 29-32; Pkc_Phospho_Site 42-44;
DEX0255_118	Ck2_Phospho_Site 7-10; Pkc_Phospho_Site 7-9; Prokar_Lipoprotein 21-31;
DEX0255_119	Myristyl 13-18; Pkc_Phospho_Site 21-23;
DEX0255_120	Asn_Glycosylation 17-20; Ck2_Phospho_Site 52-55;
DEX0255_121	Ck2_Phospho_Site 13-16; Pkc_Phospho_Site 18-20;44-46;
DEX0255_122	Myristyl 2-7;
DEX0255_123	Ck2_Phospho_Site 46-49;80-83; Myristyl 20-25;43-48;
DEX0255_125	Camp_Phospho_Site 40-43; Pkc_Phospho_Site 55-57;59-61;67-69;75-77; Tyr_Phospho_Site 68-76;69-76;
DEX0255_126	Pkc_Phospho_Site 28-30;
DEX0255_127	Ck2_Phospho_Site 36-39;75-78; Myristyl 91-96;
DEX0255_128	Asn_Glycosylation 17-20; Camp_Phospho_Site 4-7; Ck2_Phospho_Site 9-12; Pkc_Phospho_Site 9-11;
DEX0255_130	Myristyl 9-14; Pkc_Phospho_Site 14-16;41-43;45-47;
DEX0255_131	Ck2_Phospho_Site 13-16;22-25; Myristyl 17-22;
DEX0255_132	Myristyl 12-17;
DEX0255_133	Asn_Glycosylation 33-36;
DEX0255_136	Ck2_Phospho_Site 22-25; Myristyl 5-10;48-53;56-61;
DEX0255_137	Amidation 31-34; Camp_Phospho_Site 33-36; Pkc_Phospho_Site 31-33;
DEX0255_138	Pkc_Phospho_Site 28-30;

DEX0255_139	Pkc_Phospho_Site 22-24;23-25;
DEX0255_140	Asn_Glycosylation 41-44; Pkc_Phospho_Site 43-45;53-55;
DEX0255_141	Amidation 274-277;543-546;814-817; Asn_Glycosylation 64-67;294-297;391-394;610-613;611-614;660-663; Ck2_Phospho_Site 19-22;114-117;128-131;185-188;263-266;286-289;398-401;426-429;532-535;555-558;669-672;697-700;803-806;826-829;841-844; Myristyl 89-94;147-152;320-325;345-350;446-451;458-463;589-594;614-619;717-722;729-734; Pkc_Phospho_Site 75-77;76-78;102-104;148-150;274-276;381-383;459-461;543-545;650-652;730-732;814-816;
DEX0255_142	Camp_Phospho_Site 18-21;
DEX0255_143	Pkc_Phospho_Site 16-18;
DEX0255_144	Ck2_Phospho_Site 24-27;
DEX0255_145	Camp_Phospho_Site 13-16; Ck2_Phospho_Site 29-32; Myristyl 25-30;
DEX0255_147	Myristyl 43-48; Prokar_Lipoprotein 37-47;
DEX0255_148	Myristyl 9-14;11-16;32-37;
DEX0255_149	Asn_Glycosylation 5-8; Pkc_Phospho_Site 18-20;
DEX0255_150	Asn_Glycosylation 9-12; Ck2_Phospho_Site 23-26;
DEX0255_151	Myristyl 68-73;
DEX0255_152	Leucine_Zipper 52-73; Pkc_Phospho_Site 56-58;
DEX0255_154	Pkc_Phospho_Site 18-20;
DEX0255_155	Ck2_Phospho_Site 32-35; Myristyl 28-33;57-62; Pkc_Phospho_Site 45-47;
DEX0255_157	Asn_Glycosylation 16-19;28-31; Pkc_Phospho_Site 6-8;19-21;25-27;38-40;
DEX0255_159	Ck2_Phospho_Site 27-30;
DEX0255_164	Amidation 23-26; Pkc_Phospho_Site 33-35;70-72;123-125;
DEX0255_165	Ck2_Phospho_Site 2-5;42-45; Myristyl 30-35; Pkc_Phospho_Site 11-13;35-37;42-44;
DEX0255_166	Asn_Glycosylation 18-21; Myristyl 25-30; Pkc_Phospho_Site 29-31;
DEX0255_168	Myristyl 91-96; Pkc_Phospho_Site 60-62;
DEX0255_170	Ck2_Phospho_Site 22-25; Pkc_Phospho_Site 31-33;
DEX0255_171	Asn_Glycosylation 141-144; Ck2_Phospho_Site 41-44; Myristyl 138-143; Pkc_Phospho_Site 41-43;69-71;101-103;120-122;143-145;150-152; Protein_Kinase_Atp 127-149; Rgd 49-51; Tyr_Phospho_Site 59-66;
DEX0255_173	Amidation 38-41; Pkc_Phospho_Site 28-30;
DEX0255_174	Asn_Glycosylation 93-96; Ck2_Phospho_Site 16-19;46-49;87-90;
DEX0255_175	Amidation 82-85; Asn_Glycosylation 120-123;344-347;406-409; Camp_Phospho_Site 5-8;6-9;486-489; Ck2_Phospho_Site 8-11;10-13;28-31;32-35;64-67;70-73;99-102;155-158;267-270;297-300;338-341;404-407;408-411;472-475; Leucine_Zipper 357-378;364-385; Myristyl 97-102;159-164;185-190;193-198; Pkc_Phospho_Site 4-6;32-34;58-60;211-213;214-216;361-363;433-435;

DEX0255_176	Amidation 25-28; Camp_Phospho_Site 27-30;61-64; Ck2_Phospho_Site 67-70;73-76;109-112;127-130; Leucine_Zipper 93-114; Myristyl 29-34;79-84;

Example 6: Method of Determining Alterations in a Gene Corresponding to a Polynucleotide

RNA is isolated from individual patients or from a family of individuals that have a phenotype of interest. cDNA is then generated from these RNA samples using protocols known in the art. *See*, Sambrook (2001), *supra*. The cDNA is then used as a template for PCR, employing primers surrounding regions of interest in SEQ ID NO: 1 through 100. Suggested PCR conditions consist of 35 cycles at 95°C for 30 seconds; 60-120 seconds at 52-58°C; and 60-120 seconds at 70°C, using buffer solutions described in Sidransky *et al., Science* 252(5006): 706-9 (1991). *See also* Sidransky *et al., Science* 278(5340): 1054-9 (1997).

PCR products are then sequenced using primers labeled at their 5' end with T4 polynucleotide kinase, employing SequiTherm Polymerase. (Epicentre Technologies). The intron-exon borders of selected exons is also determined and genomic PCR products analyzed to confirm the results. PCR products harboring suspected mutations are then cloned and sequenced to validate the results of the direct sequencing. PCR products is cloned into T-tailed vectors as described in Holton *et al., Nucleic Acids Res.*, 19: 1156 (1991) and sequenced with T7 polymerase (United States Biochemical). Affected individuals are identified by mutations not present in unaffected individuals.

Genomic rearrangements may also be determined. Genomic clones are nick-translated with digoxigenin deoxyuridine 5' triphosphate (Boehringer Manheim), and FISH is performed as described in Johnson *et al., Methods Cell Biol.* 35: 73-99 (1991). Hybridization with the labeled probe is carried out using a vast excess of human cot-1 DNA for specific hybridization to the corresponding genomic locus.

Chromosomes are counterstained with 4,6-diamino-2-phenylidole and propidium iodide, producing a combination of C-and R-bands. Aligned images for precise mapping are obtained using a triple-band filter set (Chroma Technology, Brattleboro, VT) in combination with a cooled charge-coupled device camera (Photometrics, Tucson, AZ) and variable excitation wavelength filters. *Id.* Image collection, analysis and chromosomal fractional length measurements are performed using the ISee Graphical

Program System. (Inovision Corporation, Durham, NC.) Chromosome alterations of the genomic region hybridized by the probe are identified as insertions, deletions, and translocations. These alterations are used as a diagnostic marker for an associated disease.

Example 7: Method of Detecting Abnormal Levels of a Polypeptide in a Biological Sample

Antibody-sandwich ELISAs are used to detect polypeptides in a sample, preferably a biological sample. Wells of a microtiter plate are coated with specific antibodies, at a final concentration of 0.2 to 10 μg/ml. The antibodies are either monoclonal or polyclonal and are produced by the method described above. The wells are blocked so that non-specific binding of the polypeptide to the well is reduced. The coated wells are then incubated for > 2 hours at RT with a sample containing the polypeptide. Preferably, serial dilutions of the sample should be used to validate results. The plates are then washed three times with deionized or distilled water to remove unbound polypeptide. Next, 50 μl of specific antibody-alkaline phosphatase conjugate, at a concentration of 25-400 ng, is added and incubated for 2 hours at room temperature. The plates are again washed three times with deionized or distilled water to remove unbound conjugate. 75 μl of 4-methylumbelliferyl phosphate (MUP) or p-nitrophenyl phosphate (NPP) substrate solution are added to each well and incubated 1 hour at room temperature.

The reaction is measured by a microtiter plate reader. A standard curve is prepared, using serial dilutions of a control sample, and polypeptide concentrations are plotted on the X-axis (log scale) and fluorescence or absorbance on the Y-axis (linear scale). The concentration of the polypeptide in the sample is calculated using the standard curve.

Example 8: Formulating a Polypeptide

The secreted polypeptide composition will be formulated and dosed in a fashion consistent with good medical practice, taking into account the clinical condition of the individual patient (especially the side effects of treatment with the secreted polypeptide alone), the site of delivery, the method of administration, the scheduling of

administration, and other factors known to practitioners. The "effective amount" for purposes herein is thus determined by such considerations.

As a general proposition, the total pharmaceutically effective amount of secreted polypeptide administered parenterally per dose will be in the range of about 1 , μg/kg/day to 10 mg/kg/day of patient body weight, although, as noted above, this will be subject to therapeutic discretion. More preferably, this dose is at least 0.01 mg/kg/day, and most preferably for humans between about 0.01 and 1 mg/kg/day for the hormone. If given continuously, the secreted polypeptide is typically administered at a dose rate of about 1 μg/kg/hour to about 50 mg/kg/hour, either by 1-4 injections per day or by continuous subcutaneous infusions, for example, using a mini-pump. An intravenous bag solution may also be employed. The length of treatment needed to observe changes and the interval following treatment for responses to occur appears to vary depending on the desired effect.

Pharmaceutical compositions containing the secreted protein of the invention are administered orally, rectally, parenterally, intracistemally, intravaginally, intraperitoneally, topically (as by powders, ointments, gels, drops or transdermal patch), bucally, or as an oral or nasal spray. "Pharmaceutically acceptable carrier" refers to a non-toxic solid, semisolid or liquid filler, diluent, encapsulating material or formulation auxiliary of any type. The term "parenteral" as used herein refers to modes of administration which include intravenous, intramuscular, intraperitoneal, intrasternal, subcutaneous and intraarticular injection and infusion.

The secreted polypeptide is also suitably administered by sustained-release systems. Suitable examples of sustained-release compositions include semipermeable polymer matrices in the form of shaped articles, e. g., films, or microcapsules. Sustained-release matrices include polylactides (U. S. Pat. No.3,773,919, EP 58,481), copolymers of L-glutamic acid and gamma-ethyl-L-glutamate (Sidman, U. et al., Biopolymers 22: 547-556 (1983)), poly (2-hydroxyethyl methacrylate) (R. Langer et al., J. Biomed. Mater. Res. 15: 167-277 (1981), and R. Langer, Chem. Tech. 12: 98-105 (1982)), ethylene vinyl acetate (R. Langer et al.) or poly-D- (-)-3-hydroxybutyric acid (EP 133,988). Sustained-release compositions also include liposomally entrapped polypeptides. Liposomes containing the secreted polypeptide are prepared by methods known per se: DE Epstein et al., Proc. Natl. Acad. Sci. USA 82: 3688-3692 (1985); Hwang et al., Proc. Natl. Acad.

Sci. USA 77: 4030-4034 (1980); EP 52,322; EP 36,676; EP 88,046; EP 143,949; EP 142,641; Japanese Pat. Appl. 83-118008; U. S. Pat. Nos. 4,485,045 and 4,544,545; and EP 102,324. Ordinarily, the liposomes are of the small (about 200-800 Angstroms) unilamellar type in which the lipid content is greater than about 30 mol. percent cholesterol, the selected proportion being adjusted for the optimal secreted polypeptide therapy.

For parenteral administration, in one embodiment, the secreted polypeptide is formulated generally by mixing it at the desired degree of purity, in a unit dosage injectable form (solution, suspension, or emulsion), with a pharmaceutically acceptable carrier, I. e., one that is non-toxic to recipients at the dosages and concentrations employed and is compatible with other ingredients of the formulation.

For example, the formulation preferably does not include oxidizing agents and other compounds that are known to be deleterious to polypeptides. Generally, the formulations are prepared by contacting the polypeptide uniformly and intimately with liquid carriers or finely divided solid carriers or both. Then, if necessary, the product is shaped into the desired formulation. Preferably the carrier is a parenteral carrier, more preferably a solution that is isotonic with the blood of the recipient. Examples of such carrier vehicles include water, saline, Ringer's solution, and dextrose solution. Non-aqueous vehicles such as fixed oils and ethyl oleate are also useful herein, as well as liposomes.

The carrier suitably contains minor amounts of additives such as substances that enhance isotonicity and chemical stability. Such materials are non-toxic to recipients at the dosages and concentrations employed, and include buffers such as phosphate, citrate, succinate, acetic acid, and other organic acids or their salts; antioxidants such as ascorbic acid; low molecular weight (less than about ten residues) polypeptides, e. g., polyarginine or tripeptides; proteins, such as serum albumin, gelatin, or immunoglobulins; hydrophilic polymers such as polyvinylpyrrolidone; amino acids, such as glycine, glutamic acid, aspartic acid, or arginine; monosaccharides, disaccharides, and other carbohydrates including cellulose or its derivatives, glucose, manose, or dextrins; chelating agents such as EDTA; sugar alcohols such as mannitol or sorbitol; counterions such as sodium; and/or nonionic surfactants such as polysorbates, poloxamers, or PEG.

The secreted polypeptide is typically formulated in such vehicles at a concentration of about 0.1 mg/ml to 100 mg/ml, preferably 1-10 mg/ml, at a pH of about 3 to 8. It will be understood that the use of certain of the foregoing excipients, carriers, or stabilizers will result in the formation of polypeptide salts.

Any polypeptide to be used for therapeutic administration can be sterile. Sterility is readily accomplished by filtration through sterile filtration membranes (e. g., 0.2 micron membranes). Therapeutic polypeptide compositions generally are placed into a container having a sterile access port, for example, an intravenous solution bag or vial having a stopper pierceable by a hypodermic injection needle.

Polypeptides ordinarily will be stored in unit or multi-dose containers, for example, sealed ampules or vials, as an aqueous solution or as a lyophilized formulation for reconstitution. As an example of a lyophilized formulation, 10-ml vials are filled with 5 ml of sterile-filtered 1 % (w/v) aqueous polypeptide solution, and the resulting mixture is lyophilized. The infusion solution is prepared by reconstituting the lyophilized polypeptide using bacteriostatic Water-for-Injection.

The invention also provides a pharmaceutical pack or kit comprising one or more containers filled with one or more of the ingredients of the pharmaceutical compositions of the invention. Associated with such container (s) can be a notice in the form prescribed by a governmental agency regulating the manufacture, use or sale of pharmaceuticals or biological products, which notice reflects approval by the agency of manufacture, use or sale for human administration. In addition, the polypeptides of the present invention may be employed in conjunction with other therapeutic compounds.

Example 9: Method of Treating Decreased Levels of the Polypeptide

It will be appreciated that conditions caused by a decrease in the standard or normal expression level of a secreted protein in an individual can be treated by administering the polypeptide of the present invention, preferably in the secreted form. Thus, the invention also provides a method of treatment of an individual in need of an increased level of the polypeptide comprising administering to such an individual a pharmaceutical composition comprising an amount of the polypeptide to increase the activity level of the polypeptide in such an individual.

For example, a patient with decreased levels of a polypeptide receives a daily dose 0.1-100 μg/kg of the polypeptide for six consecutive days. Preferably, the

polypeptide is in the secreted form. The exact details of the dosing scheme, based on administration and formulation, are provided above.

Example 10: Method of Treating Increased Levels of the Polypeptide

Antisense technology is used to inhibit production of a polypeptide of the present invention. This technology is one example of a method of decreasing levels of a polypeptide, preferably a secreted form, due to a variety of etiologies, such as cancer.

For example, a patient diagnosed with abnormally increased levels of a polypeptide is administered intravenously antisense polynucleotides at 0.5, 1.0, 1.5, 2.0 and 3.0 mg/kg day for 21 days. This treatment is repeated after a 7-day rest period if the treatment was well tolerated. The formulation of the antisense polynucleotide is provided above.

Example 11: Method of Treatment Using Gene Therapy

One method of gene therapy transplants fibroblasts, which are capable of expressing a polypeptide, onto a patient. Generally, fibroblasts are obtained from a subject by skin biopsy. The resulting tissue is placed in tissue-culture medium and separated into small pieces. Small chunks of the tissue are placed on a wet surface of a tissue culture flask, approximately ten pieces are placed in each flask. The flask is turned upside down, closed tight and left at room temperature over night. After 24 hours at room temperature, the flask is inverted and the chunks of tissue remain fixed to the bottom of the flask and fresh media (e. g., Ham's F12 media, with 10% FBS, penicillin and streptomycin) is added. The flasks are then incubated at 37°C for approximately one week.

At this time, fresh media is added and subsequently changed every several days. After an additional two weeks in culture, a monolayer of fibroblasts emerge. The monolayer is trypsinized and scaled into larger flasks. pMV-7 (Kirschmeier, P. T. et al., DNA, 7: 219-25 (1988)), flanked by the long terminal repeats of the Moloney murine sarcoma virus, is digested with EcoRI and HindIII and subsequently treated with calf intestinal phosphatase. The linear vector is fractionated on agarose gel and purified, using glass beads.

The cDNA encoding a polypeptide of the present invention can be amplified using PCR primers which correspond to the 5'and 3'end sequences respectively as set forth in Example 1. Preferably, the 5'primer contains an EcoRI site and the 3'primer

includes a HindIII site. Equal quantities of the Moloney murine sarcoma virus linear backbone and the amplified EcoRI and HindIII fragment are added together, in the presence of T4 DNA ligase. The resulting mixture is maintained under conditions appropriate for ligation of the two fragments. The ligation mixture is then used to transform bacteria HB 101, which are then plated onto agar containing kanamycin for the purpose of confirming that the vector has the gene of interest properly inserted.

The amphotropic pA317 or GP+aml2 packaging cells are grown in tissue culture to confluent density in Dulbecco's Modified Eagles Medium (DMEM) with 10% calf serum (CS), penicillin and streptomycin. The MSV vector containing the gene is then added to the media and the packaging cells transduced with the vector. The packaging cells now produce infectious viral particles containing the gene (the packaging cells are now referred to as producer cells).

Fresh media is added to the transduced producer cells, and subsequently, the media is harvested from a 10 cm plate of confluent producer cells. The spent media, containing the infectious viral particles, is filtered through a millipore filter to remove detached producer cells and this media is then used to infect fibroblast cells. Media is removed from a sub-confluent plate of fibroblasts and quickly replaced with the media from the producer cells. This media is removed and replaced with fresh media.

If the titer of virus is high, then virtually all fibroblasts will be infected and no selection is required. If the titer is very low, then it is necessary to use a retroviral vector that has a selectable marker, such as neo or his. Once the fibroblasts have been efficiently infected, the fibroblasts are analyzed to determine whether protein is produced.

The engineered fibroblasts are then transplanted onto the host, either alone or after having been grown to confluence on cytodex 3 microcarrier beads.

Example 12: Method of Treatment Using Gene Therapy-*In Vivo*

Another aspect of the present invention is using *in vivo* gene therapy methods to treat disorders, diseases and conditions. The gene therapy method relates to the introduction of naked nucleic acid (DNA, RNA, and antisense DNA or RNA) sequences into an animal to increase or decrease the expression of the polypeptide.

The polynucleotide of the present invention may be operatively linked to a promoter or any other genetic elements necessary for the expression of the polypeptide by the target tissue. Such gene therapy and delivery techniques and methods are known

in the art, see, for example, W0 90/11092, W0 98/11779; U. S. Patent 5,693,622; 5,705,151; 5,580,859; Tabata H. et al. (1997) Cardiovasc. Res. 35 (3): 470-479, Chao J et al. (1997) Pharmacol. Res. 35 (6): 517-522, Wolff J. A. (1997) Neuromuscul. Disord. 7 (5): 314-318, Schwartz B. et al. (1996) Gene Ther. 3 (5): 405-411, Tsurumi Y. et al. (1996) Circulation 94 (12): 3281-3290 (incorporated herein by reference).

The polynucleotide constructs may be delivered by any method that delivers injectable materials to the cells of an animal, such as, injection into the interstitial space of tissues (heart, muscle, skin, lung, liver, intestine and the like). The polynucleotide constructs can be delivered in a pharmaceutically acceptable liquid or aqueous carrier.

The term "naked" polynucleotide, DNA or RNA, refers to sequences that are free from any delivery vehicle that acts to assist, promote, or facilitate entry into the cell, including viral sequences, viral particles, liposome formulations, lipofectin or precipitating agents and the like. However, the polynucleotides of the present invention may also be delivered in liposome formulations (such as those taught in Felgner P. L. et al. (1995) Ann. NY Acad. Sci. 772: 126-139 and Abdallah B. et al. (1995) Biol. Cell 85 (1): 1-7) which can be prepared by methods well known to those skilled in the art.

The polynucleotide vector constructs used in the gene therapy method are preferably constructs that will not integrate into the host genome nor will they contain sequences that allow for replication. Any strong promoter known to those skilled in the art can be used for driving the expression of DNA. Unlike other gene therapies techniques, one major advantage of introducing naked nucleic acid sequences into target cells is the transitory nature of the polynucleotide synthesis in the cells. Studies have shown that non-replicating DNA sequences can be introduced into cells to provide production of the desired polypeptide for periods of up to six months.

The polynucleotide construct can be delivered to the interstitial space of tissues within the an animal, including of muscle, skin, brain, lung, liver, spleen, bone marrow, thymus, heart, lymph, blood, bone, cartilage, pancreas, kidney, gall bladder, stomach, intestine, testis, ovary, uterus, rectum, nervous system, eye, gland, and connective tissue. Interstitial space of the tissues comprises the intercellular fluid, mucopolysaccharide matrix among the reticular fibers of organ tissues, elastic fibers in the walls of vessels or chambers, collagen fibers of fibrous tissues, or that same matrix within connective tissue ensheathing muscle cells or in the lacunae of bone. It is similarly the space occupied by

the plasma of the circulation and the lymph fluid of the lymphatic channels. Delivery to the interstitial space of muscle tissue is preferred for the reasons discussed below. They may be conveniently delivered by injection into the tissues comprising these cells. They are preferably delivered to and expressed in persistent, non-dividing cells which are differentiated, although delivery and expression may be achieved in non-differentiated or less completely differentiated cells, such as, for example, stem cells of blood or skin fibroblasts. *In vivo* muscle cells are particularly competent in their ability to take up and express polynucleotides.

For the naked polynucleotide injection, an effective dosage amount of DNA or RNA will be in the range of from about 0.05 μg/kg body weight to about 50 mg/kg body weight. Preferably the dosage will be from about 0.005 mg/kg to about 20 mg/kg and more preferably from about 0.05 mg/kg to about 5 mg/kg. Of course, as the artisan of ordinary skill will appreciate, this dosage will vary according to the tissue site of injection. The appropriate and effective dosage of nucleic acid sequence can readily be determined by those of ordinary skill in the art and may depend on the condition being treated and the route of administration. The preferred route of administration is by the parenteral route of injection into the interstitial space of tissues. However, other parenteral routes may also be used, such as, inhalation of an aerosol formulation particularly for delivery to lungs or bronchial tissues, throat or mucous membranes of the nose. In addition, naked polynucleotide constructs can be delivered to arteries during angioplasty by the catheter used in the procedure.

The dose response effects of injected polynucleotide in muscle *in vivo* is determined as follows. Suitable template DNA for production of mRNA coding for polypeptide of the present invention is prepared in accordance with a standard recombinant DNA methodology. The template DNA, which may be either circular or linear, is either used as naked DNA or complexed with liposomes. The quadriceps muscles of mice are then injected with various amounts of the template DNA.

Five to six week old female and male Balb/C mice are anesthetized by intraperitoneal injection with 0.3 ml of 2.5% Avertin. A 1.5 cm incision is made on the anterior thigh, and the quadriceps muscle is directly visualized. The template DNA is injected in 0.1 ml of carrier in a 1 cc syringe through a 27 gauge needle over one minute, approximately 0.5 cm from the distal insertion site of the muscle into the knee and about

0.2 cm deep. A suture is placed over the injection site for future localization, and the skin is closed with stainless steel clips.

After an appropriate incubation time (e. g., 7 days) muscle extracts are prepared by excising the entire quadriceps. Every fifth 15 um cross-section of the individual quadriceps muscles is histochemically stained for protein expression. A time course for protein expression may be done in a similar fashion except that quadriceps from different mice are harvested at different times. Persistence of DNA in muscle following injection may be determined by Southern blot analysis after preparing total cellular DNA and HIRT supernatants from injected and control mice.

The results of the above experimentation in mice can be use to extrapolate proper dosages and other treatment parameters in humans and other animals using naked DNA.

Example 13: Transgenic Animals

The polypeptides of the invention can also be expressed in transgenic animals. Animals of any species, including, but not limited to, mice, rats, rabbits, hamsters, guinea pigs, pigs, micro-pigs, goats, sheep, cows and non-human primates, e. g., baboons, monkeys, and chimpanzees may be used to generate transgenic animals. In a specific embodiment, techniques described herein or otherwise known in the art, are used to express polypeptides of the invention in humans, as part of a gene therapy protocol.

Any technique known in the art may be used to introduce the transgene (i. e., polynucleotides of the invention) into animals to produce the founder lines of transgenic animals. Such techniques include, but are not limited to, pronuclear microinjection (Paterson et al., Appl. Microbiol. Biotechnol. 40: 691-698 (1994); Carver et al., Biotechnology (NY) 11: 1263-1270 (1993); Wright et al., Biotechnology (NY) 9: 830-834 (1991); and Hoppe et al., U. S. Patent 4,873,191 (1989)); retrovirus mediated gene transfer into germ lines (Van der Putten et al., Proc. Natl. Acad. Sci., USA 82: 6148-6152 (1985)), blastocysts or embryos; gene targeting in embryonic stem cells (Thompson et al., Cell 56: 313-321 (1989)); electroporation of cells or embryos (Lo, 1983, Mol Cell. Biol. 3: 1803-1814 (1983)); introduction of the polynucleotides of the invention using a gene gun (see, e. g., Ulmer et al., Science 259: 1745 (1993); introducing nucleic acid constructs into embryonic pleuripotent stem cells and transferring the stem cells back into the blastocyst; and sperm mediated gene transfer (Lavitrano et al., Cell 57: 717-723 (1989); etc. For a review of such techniques, see Gordon,"Transgenic Animals," Intl.

Rev. Cytol. 115: 171-229 (1989), which is incorporated by reference herein in its entirety.

Any technique known in the art may be used to produce transgenic clones containing polynucleotides of the invention, for example, nuclear transfer into enucleated oocytes of nuclei from cultured embryonic, fetal, or adult cells induced to quiescence (Campell et al., Nature 380: 64-66 (1996); Wilmut et al., Nature 385: 810813 (1997)).

The present invention provides for transgenic animals that carry the transgene in all their cells, as well as animals which carry the transgene in some, but not all their cells, I. e., mosaic animals or chimeric. The transgene may be integrated as a single transgene or as multiple copies such as in concatamers, e. g., head-to-head tandems or head-to-tail tandems. The transgene may also be selectively introduced into and activated in a particular cell type by following, for example, the teaching of Lasko et al. (Lasko et al., Proc. Natl. Acad. Sci. USA 89: 6232-6236 (1992)). The regulatory sequences required for such a cell-type specific activation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art. When it is desired that the polynucleotide transgene be integrated into the chromosomal site of the endogenous gene, gene targeting is preferred. Briefly, when such a technique is to be utilized, vectors containing some nucleotide sequences homologous to the endogenous gene are designed for the purpose of integrating, via homologous recombination with chromosomal sequences, into and disrupting the function of the nucleotide sequence of the endogenous gene. The transgene may also be selectively introduced into a particular cell type, thus inactivating the endogenous gene in only that cell type, by following, for example, the teaching of Gu et al. (Gu et al., Science 265: 103-106 (1994)). The regulatory sequences required for such a cell-type specific inactivation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art.

Once transgenic animals have been generated, the expression of the recombinant gene may be assayed utilizing standard techniques. Initial screening may be accomplished by Southern blot analysis or PCR techniques to analyze animal tissues to verify that integration of the transgene has taken place. The level of mRNA expression of the transgene in the tissues of the transgenic animals may also be assessed using techniques which include, but are not limited to, Northern blot analysis of tissue samples obtained from the animal, in situ hybridization analysis, and reverse transcriptase-PCR

(rt-PCR). Samples of transgenic gene-expressing tissue may also be evaluated immunocytochemically or immunohistochemically using antibodies specific for the transgene product.

Once the founder animals are produced, they may be bred, inbred, outbred, or crossbred to produce colonies of the particular animal. Examples of such breeding strategies include, but are not limited to: outbreeding of founder animals with more than one integration site in order to establish separate lines; inbreeding of separate lines in order to produce compound transgenics that express the transgene at higher levels because of the effects of additive expression of each transgene; crossing of heterozygous transgenic animals to produce animals homozygous for a given integration site in order to both augment expression and eliminate the need for screening of animals by DNA analysis; crossing of separate homozygous lines to produce compound heterozygous or homozygous lines; and breeding to place the transgene on a distinct background that is appropriate for an experimental model of interest.

Transgenic animals of the invention have uses which include, but are not limited to, animal model systems useful in elaborating the biological function of polypeptides of the present invention, studying conditions and/or disorders associated with aberrant expression, and in screening for compounds effective in ameliorating such conditions and/or disorders.

Example 14: Knock-Out Animals

Endogenous gene expression can also be reduced by inactivating or "knocking out" the gene and/or its promoter using targeted homologous recombination. (E. g., see Smithies et al., Nature 317: 230-234 (1985); Thomas & Capecchi, Cell 51: 503512 (1987); Thompson et al., Cell 5: 313-321 (1989); each of which is incorporated by reference herein in its entirety). For example, a mutant, non-functional polynucleotide of the invention (or a completely unrelated DNA sequence) flanked by DNA homologous to the endogenous polynucleotide sequence (either the coding regions or regulatory regions of the gene) can be used, with or without a selectable marker and/or a negative selectable marker, to transfect cells that express polypeptides of the invention *in vivo*. In another embodiment, techniques known in the art are used to generate knockouts in cells that contain, but do not express the gene of interest. Insertion of the DNA construct, via targeted homologous recombination, results in inactivation of the targeted gene. Such

approaches are particularly suited in research and agricultural fields where modifications to embryonic stem cells can be used to generate animal offspring with an inactive targeted gene (e. g., see Thomas & Capecchi 1987 and Thompson 1989, supra). However this approach can be routinely adapted for use in humans provided the recombinant DNA constructs are directly administered or targeted to the required site *in vivo* using appropriate viral vectors that will be apparent to those of skill in the art.

In further embodiments of the invention, cells that are genetically engineered to express the polypeptides of the invention, or alternatively, that are genetically engineered not to express the polypeptides of the invention (e. g., knockouts) are administered to a patient *in vivo*. Such cells may be obtained from the patient (I. e., animal, including human) or an MHC compatible donor and can include, but are not limited to fibroblasts, bone marrow cells, blood cells (e. g., lymphocytes), adipocytes, muscle cells, endothelial cells etc. The cells are genetically engineered *in vitro* using recombinant DNA techniques to introduce the coding sequence of polypeptides of the invention into the cells, or alternatively, to disrupt the coding sequence and/or endogenous regulatory sequence associated with the polypeptides of the invention, e. g., by transduction (using viral vectors, and preferably vectors that integrate the transgene into the cell genome) or transfection procedures, including, but not limited to, the use of plasmids, cosmids, YACs, naked DNA, electroporation, liposomes, etc.

The coding sequence of the polypeptides of the invention can be placed under the control of a strong constitutive or inducible promoter or promoter/enhancer to achieve expression, and preferably secretion, of the polypeptides of the invention. The engineered cells which express and preferably secrete the polypeptides of the invention can be introduced into the patient systemically, e. g., in the circulation, or intraperitoneally.

Alternatively, the cells can be incorporated into a matrix and implanted in the body, e. g., genetically engineered fibroblasts can be implanted as part of a skin graft; genetically engineered endothelial cells can be implanted as part of a lymphatic or vascular graft. (See, for example, Anderson et al. U. S. Patent 5,399,349; and Mulligan & Wilson, U. S. Patent 5,460,959 each of which is incorporated by reference herein in its entirety).

When the cells to be administered are non-autologous or non-MHC compatible cells, they can be administered using well known techniques which prevent the

development of a host immune response against the introduced cells. For example, the cells may be introduced in an encapsulated form which, while allowing for an exchange of components with the immediate extracellular environment, does not allow the introduced cells to be recognized by the host immune system.

Transgenic and "knock-out" animals of the invention have uses which include, but are not limited to, animal model systems useful in elaborating the biological function of polypeptides of the present invention, studying conditions and/or disorders associated with aberrant expression, and in screening for compounds effective in ameliorating such conditions and/or disorders.

All patents, patent publications, and other published references mentioned herein are hereby incorporated by reference in their entireties as if each had been individually and specifically incorporated by reference herein. While preferred illustrative embodiments of the present invention are described, one skilled in the art will appreciate that the present invention can be practiced by other than the described embodiments, which are presented for purposes of illustration only and not by way of limitation. The present invention is limited only by the claims that follow.

CLAIMS

We claim:

1. An isolated nucleic acid molecule comprising

(a) a nucleic acid molecule comprising a nucleic acid sequence that encodes an amino acid sequence of SEQ ID NO: 101 through 176;

(b) a nucleic acid molecule comprising a nucleic acid sequence of SEQ ID NO: 1 through 100;

(c) a nucleic acid molecule that selectively hybridizes to the nucleic acid molecule of (a) or (b); or

(d) a nucleic acid molecule having at least 60% sequence identity to the nucleic acid molecule of (a) or (b).

2. The nucleic acid molecule according to claim 1, wherein the nucleic acid molecule is a cDNA.

3. The nucleic acid molecule according to claim 1, wherein the nucleic acid molecule is genomic DNA.

4. The nucleic acid molecule according to claim 1, wherein the nucleic acid molecule is a mammalian nucleic acid molecule.

5. The nucleic acid molecule according to claim 4, wherein the nucleic acid molecule is a human nucleic acid molecule.

6. A method for determining the presence of a colon specific nucleic acid (CSNA) in a sample, comprising the steps of:

(a) contacting the sample with the nucleic acid molecule according to claim 1 under conditions in which the nucleic acid molecule will selectively hybridize to a colon specific nucleic acid; and

(b) detecting hybridization of the nucleic acid molecule to a CSNA in the sample, wherein the detection of the hybridization indicates the presence of a CSNA in the sample.

7. A vector comprising the nucleic acid molecule of claim 1.

8. A host cell comprising the vector according to claim 7.

9. A method for producing a polypeptide encoded by the nucleic acid molecule according to claim 1, comprising the steps of (a) providing a host cell comprising the nucleic acid molecule operably linked to one or more expression control sequences, and (b) incubating the host cell under conditions in which the polypeptide is produced.

10. A polypeptide encoded by the nucleic acid molecule according to claim 1.

11. An isolated polypeptide selected from the group consisting of:

(a) a polypeptide comprising an amino acid sequence with at least 60% sequence identity to of SEQ ID NO: 101 through 176; or

(b) a polypeptide comprising an amino acid sequence encoded by a nucleic acid molecule comprising a nucleic acid sequence of SEQ ID NO: 1 through 100.

12. An antibody or fragment thereof that specifically binds to the polypeptide according to claim 11.

13. A method for determining the presence of a colon specific protein in a sample, comprising the steps of:

(a) contacting the sample with the antibody according to claim 12 under conditions in which the antibody will selectively bind to the colon specific protein; and

(b) detecting binding of the antibody to a colon specific protein in the sample, wherein the detection of binding indicates the presence of a colon specific protein in the sample.

14. A method for diagnosing and monitoring the presence and metastases of colon cancer in a patient, comprising the steps of:

(a) determining an amount of the nucleic acid molecule of claim 1 or a polypeptide of claim 6 in a sample of a patient; and

(b) comparing the amount of the determined nucleic acid molecule or the polypeptide in the sample of the patient to the amount of the colon specific marker in a normal control; wherein a difference in the amount of the nucleic acid molecule or the polypeptide in the sample compared to the amount of the nucleic acid molecule or the polypeptide in the normal control is associated with the presence of colon cancer.

15. A kit for detecting a risk of cancer or presence of cancer in a patient, said kit comprising a means for determining the presence the nucleic acid molecule of claim 1 or a polypeptide of claim 6 in a sample of a patient.

16. A method of treating a patient with colon cancer, comprising the step of administering a composition according to claim 12 to a patient in need thereof, wherein said administration induces an immune response against the colon cancer cell expressing the nucleic acid molecule or polypeptide.

17. A vaccine comprising the polypeptide or the nucleic acid encoding the polypeptide of claim 11.

SEQUENCE LISTING

<110> Sun, Yongming
Recipon, Herve
Ghosh, Malavika
Liu, Chenghua
diaDexus, Inc.

<120> Compositions and Methods Relating to Colon Specific
Genes and Proteins

<130> DEX-0256

<140>
<141>

<150> 60/244,758
<151> 2000-10-31

<160> 176

<170> PatentIn Ver. 2.1

<210> 1
<211> 108
<212> DNA
<213> Homo sapiens

<400> 1
aatctcctta gatgctacta catgacattg attggcatct gatgtctatc tgattatcag 60
gaatttacag tttcagtttc caaagaacta tatgaaaaaa ctattata 108

<210> 2
<211> 295
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (88)
<223> a, c, g or t

<220>
<221> unsure
<222> (128)
<223> a, c, g or t

<220>
<221> unsure
<222> (134)
<223> a, c, g or t

<220>
<221> unsure
<222> (171)
<223> a, c, g or t

<220>

<221> unsure
<222> (177)
<223> a, c, g or t

<220>
<221> unsure
<222> (186)
<223> a, c, g or t

<220>
<221> unsure
<222> (189)
<223> a, c, g or t

<220>
<221> unsure
<222> (195)
<223> a, c, g or t

<220>
<221> unsure
<222> (227)
<223> a, c, g or t

<220>
<221> unsure
<222> (248)
<223> a, c, g or t

<220>
<221> unsure
<222> (255)
<223> a, c, g or t

<220>
<221> unsure
<222> (290)
<223> a, c, g or t

<400> 2
ttcattccat atgtgagcat gcttaagctc cctggaagaa ccatctgaaa tgtagtaaaa 60
caattaatga gaaaagaaca ggattganat tagctgtagg atagctcaga ggaaagaaga 120
ctcactantt cagnatgtac gagtactagc taaagtaagc agagaagaac ncacttnatc 180
ttcttnganc actgngnctt tttactggtc ccttacccct accctgnata gttacattgg 240
cctctctngg tggantaatt tcccttaaga aaacagaagc tattttaacn aagcc 295

<210> 3
<211> 153
<212> DNA
<213> Homo sapiens

<400> 3
aagaaaatga aaaggcaagt aaaagttggg aaaagtatta ataataaata tgtccagtaa 60
aggacttgca tgcaaaacac ataatgaact tctacaataa tagagagaca gacaactcaa 120
aaataggcaa aagttttaag tgtacatgaa gcg 153

<210> 4

<211> 236
<212> DNA
<213> Homo sapiens

<400> 4
gaaagatgct ttatgaatct ttatatgaat cttccttctg ctgttaggtt tagtagagca 60
accccccttaa tatcactatt tctggcgctt tgattagttg cttgttggaa ttgtcaaaat 120
atctaaaaat tttaataaaa tagaacctta aagggagaaa aatcacaccg tgagcccaag 180
agaactccaa acactcacca acatggggac caaggcacct gattattatg tattga 236

<210> 5
<211> 719
<212> DNA
<213> Homo sapiens

<400> 5
gcaacaacct gagggaactg atttataatc agttgctaag gtggaaaact tgactttaat 60
aatgtatatg cagaaagaaa tgctttgctt agttgctttc tgcttatctt taagactttc 120
acactcttct tgttttttttg ttgtcattat taatataccg ttttgaaata cagaaggtct 180
aaattactaa cactacagat agcagttgtg gcaaataaaa ataagacaga gtgaaactct 240
taatcacata gctgttttct gagatttaaa ttgccactgt cactgttgag tctgatttta 300
aaatggtttc agcagttctc tctacttctc tctttccatc cttgctcttt attaaacagt 360
atttgatgcc tataatatga taacccttca acttaatttt tggaagggta cagaacaaat 420
gaggaaatct ttgtttgtaa tgtataaccc atttgttacc tgcaactgaa agacaatgaa 480
attgattatt atccaggcat gactgtggct tatttaaatc agatactttc tcttttcagc 540
tgctgaagat acttgcaatc acctgtggct tttgtattgg tagatacatc caagaaaata 600
ttgacttgaa aaaaattgtg cttagagccc tgatctggac attttaaatt cacttaaatt 660
gttttaaaaa ttgtgatata ttaatgaaat ttttattttc taactaatta tgaacctttt 719

<210> 6
<211> 687
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (185)..(330)
<223> a, c, g or t

<400> 6
gtccaaatta ccacccactc ctgagggacc caggggacca gaagtcctcg tcagtccatc 60
acagcccatt ctgcccctct gctgtcctca ccagcctcgt ggagacccaa aatggtcttc 120
ttctaggaga ggctaggggg aagggcagag gagctaggtt actagcacag agctcaagcc 180
tgatnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 240
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 300
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nggagatgga gtttataagg cacccagacc 360
agtgctggat gtccaactca acttggcaat gacgggggac acatctcctt gtttattttt 420
atgtttattt atttttttttt aatggagtct tgctctggtt gcactccagc ctgggtgaca 480
cgacgtgaga ctctgtctca aaaacagact taaaaccctt taaaagtgtg tgataaacat 540
cagcttgaca aatatgaacc cagcacccgt gaagcactgg cactatgctg ggtgctgagg 600
gagatggcat gaacagaaaa gagagacaca ggcccaagga gggaagtcca gtgtttggtg 660
ggagatgggg atgttaacaa tcacaca 687

<210> 7
<211> 475
<212> DNA

<213> Homo sapiens

<400> 7
ttttggtgag ccagccagga gtgaagacga caggtttgct gtctcctttg cctgtgggtc 60
tggggcccca gcctggggga gacccgtgac tccaggcatg ctatcaggac aacttcaacc 120
cgaggggaga tcagctctcc cgcaacctgg tgccgctctc tgacagcaca acagaaccta 180
aggggctaca ggatgattca aggaacagtg tgctacagga cctcggtatt cctgctgaga 240
gtcatttcct gtgctggatt taccatcaca gggaccagat ttatagaaaa atgaacatgt 300
attgttgcta aatggagttg gtcatggagt ggaaactaac aatatgttca cctaaatgtg 360
ctacaaccac tcaaggtctg caaacagatt cctatttaga tgttgttgag agctgaattt 420
ctggaacaag agatatagat taagactgtt gttacatgtt gcctacatga ctttc 475

<210> 8
<211> 521
<212> DNA
<213> Homo sapiens

<400> 8
ttttggtgag ccagccagga gtgaagacga caggtttgct gtctcctttg cctgtgggtc 60
tggggcccca gcctggggga gacccgtgac tccaggcatg ctatcaggac aacttcaacc 120
cgaggggaga tcagctctcc cgcaacctgg tgccgctctc tgacagcaca acagaaccta 180
aggggctaca ggatgattca aggaacagtg tgctacagga cctcggtatt cctgctgaga 240
gtcatttcct gtgctggatt taccatcaca gggaccagat ttatagaaaa atgaacatgt 300
attgttgcta aatggagttg gtcatggagt ggaaactaac aatatgttca cctaaatgtg 360
ctacaaccac tcaaggtctg caaacagatt cctatttaga tgttgttgag agctgaattt 420
ctggaacaag agatatagat taagactgtt gttacatgtt gcctacatga cttttcaaaa 480
gcttgattcc cagtggcgct ctcaacagca gtcataaaac a 521

<210> 9
<211> 743
<212> DNA
<213> Homo sapiens

<400> 9
gaaatttatg aggctacaag aatggtgaaa tgagatcagt cagtcaagtg tggaggaaag 60
taattacggc tggaaaagat ttaaactatt agaatagaga agactggatt agagaatatt 120
aatattctag aaaataacaa gcttatgaca ggaatactat atcagagtca agagaaaaca 180
aaagtatagg taaagactga atttatcata gcccaagaca aggagcagga actgacattt 240
agtcaacagt atttattgag tccctactat gtgtgccaaa ataaaaaatt aactaagtac 300
tgggtaacag ataaagaaat gattatgaaa gctggcaaat aggaagtgag agatgaaact 360
aaagatttag ttttacttgc acactgacct tgcaatttcc ccaatgtaat tttgtctgat 420
ctgagttgtg gcaagatatc ctgtttttg tacttacagg tattcctgct gagagtcatt 480
tcctgtgctg gatttaccat cacagggacc agatttatag aaaaatgaac atgtattgtt 540
gctaaatgga gttggtcatg gagtggaaac taacaatatg ttcacctgta agtttttaag 600
tgctgtgcca agaaagcctc tttggaggcc attgccataa tctactgttt acatttgtgc 660
atttagtttc cgggaaatat tttttaaacc tatcctagga gctattcaga accatagagg 720
ctatttcagt tcccctatcg act 743

<210> 10
<211> 548
<212> DNA
<213> Homo sapiens

<400> 10
atgtaaaggc ttaaatcagc acctggacta catatataag gtgatcaata aatatctact 60
ttggtatctg aatgcttcaa gatagccttt tgttcccttc taaaaaagag tgggtagtga 120

ttgttaagac ttacccgctc ataaatgttt tattacatac ttaatatgtg aagctatcaa 180
gtaaggaggt aaggaagtat gttaataggg aagaagatga taaaataccc aaataggaaa 240
ccataatgcc aactttctaa gcttaagctg actgtaacaa agaactggct ttgtaacaaa 300
gaacaagtct tactggcttc atgcaatgca aacatttttt tcttacgtat ctgacagatc 360
aaggtggtca ggtagctctc ctgagcagtt ttccaccatg tggtgattca gggatccaag 420
cacattctat cacgaggctt agccatattg gagtttttca cttcggtgat gaggatgagg 480
gagaaagtgg gagagagtga aaagagggag aaaggtggaa aagggagaaa gaaggagaag 540
gtggaaag 548

<210> 11
<211> 797
<212> DNA
<213> Homo sapiens

<400> 11
actggcttag cacttacttc ttggtgtgaa gagctagttc ttccctagag agcccaaact 60
ctggagccag gtggattgag tgggaatcct agctctgcta aagggacgta cttgcctaag 120
tctcctcctc agtaaatagg atactagtga taacatcttc atagagttgt catgagaatt 180
aaatgaataa tatatgtaaa ggcttaaatc agcacctgga ctacatatat aaggtgatca 240
ataaatatct actttggtat ctgaatgctt caagatagcc ttttgttccc ttctaaaaaa 300
gagtgggtag tgattgttaa gacttacccg ctcataaatg ttttattaca tacttaatat 360
gtgaagctat caagtaagga ggtaaggaag tatgttaata gggaagaaga tgataaaata 420
cccaaatagg aaaccataat gccaactttc taagcttaag ctgactgtaa caaagaactg 480
gctttgtaac aaagaacaag tcttactggc ttcatgcaat gcaaacattt ttttcttacg 540
tatctgacag atcaaggtgg tcaggtagct ctcctgagca gttttccacc atgtggtgat 600
tcagggatcc aagcacattc tatcacgagg cttagccata ttggagtttt tcacttcggt 660
gatgaggatg agggagaaag tgggagagag tgaaaagagg gagaaaggtg gaaaagggag 720
aaagaaggag aaggtggaaa ggagagagag agaaattgag aaagtgagaa aaaaaaaaaa 780
aaaaaacaaa aaaaaaa 797

<210> 12
<211> 558
<212> DNA
<213> Homo sapiens

<400> 12
ccatatcttg tttttgacca gtgaagggga aagaaagaaa gtaattctag aacaacctac 60
tacttacttt agagcttcac ttttatatca tctatatttg acttggaaac catctgtgtt 120
tttgtctttg tggtcagcca ttattgtttg ttttgccatg atatatgtcg gaatgatatt 180
atttgactct cctttaagta cctcctgctt ctttgagtaa ggagatccta ccagtatcta 240
acagaaaatc ttgagcttgc agaataacag aaaagaaaaa aaaaacaaaa acacaacttg 300
ccaaaacact gaaactccct ctacttatga aataaacaaa ctggcttaaa ttgggtggaa 360
tcattatggc caaatggtgt ttgggcagaa tcaccttgct gaggtcaccg tctgaatttt 420
cactgcctgt ttcatcacaa ctccccttca gtttgcccat gacatccatg aggaggcaat 480
gagcttgaga gtaacgtttc agtataaaat ggctttctct taaaaacctg atgtcatagt 540
atgggctact agcacttc 558

<210> 13
<211> 596
<212> DNA
<213> Homo sapiens

<400> 13
aaagatgaca tggattagct tcaagtaaca cataatgtgg ctggagtcag ctgatcttta 60
tgctgaaagt gtcaacagta gtgacaaata cttcaagtaa cgggtcaaaa gtctaagaca 120
gtcattctgc cagaaatggt ctgggacttc cccatacatg ggacacgtag atcaactttc 180

```
tccaagaacc accaacctgg catgcagtga tgacctctgc agtagacagg gatttaggct 240
tgattgttgt tcatctcttt ggagacataa ccctaattgt gaacttctaa attaatggcc 300
tgacaattag atcagcagct aagataaatt tcagtttgca gccccaaaga agatgttctt 360
aattagacag ttaatcattt tcaatatgga agcctaaaca cgtagaatgt ggcaataacc 420
cggaactctc agttctgaca attgagtgaa gtaatcaccc cagattcagg ttcttcattg 480
gctgacaatg agatcaaaca cccacaccag cccagtgaac accatgaggt gtcatcttcc 540
tggcccatta gtgaaccagg aacagatatt tagaaaatat tcagtaaatt ggggag     596
```

<210> 14
<211> 1114
<212> DNA
<213> Homo sapiens

<400> 14

```
gtagacattt aagttttaaa tcctgtaaag agaagcagca gagatgcaac ataggctcag 60
actttaagca gctctaattt catggctttc ctgggcctcc accatattgc catagggcct 120
gtcacatgct tatgcatcct actggctgca gcagtgcatc ctgaagacag ctcagtatag 180
ccatgggcta cttttctagt catgggctgt ggaacctgga acatgtgacc agaataccac 240
tgaccctgtg gaccagcaag cctaagttaa aggcagccat catgaggatg aaaaggagac 300
cctcaccagc cagaagggga tgtatggggt aagcagaacc ctattagaga gaaaagccaa 360
caggaatgac tgggaaaaca gcaggagggg acaaaaattg gtgtgccttg gtgacacaaa 420
aaactctggg atgacggagt tgatgacaat cgaggccaaa atcatcttca acatacccta 480
gttctaaagt atgcactatt ctccccctca gcacagccct ttgggccttt tacacaatgg 540
tgacttctaa tttggtccct actcagattt tagaactccc aagtaatgtt cccaaatgga 600
tcatattcac gaggagtagt accagtgaat taactgggtc tcaaaataca attgcacatt 660
tcagacactg ctttaaagtt gtgtcctctc attttctcat tcctgttcag ttataatagg 720
cagtcttcct cttattcacc aggggattcc atccctctga ttcaggacca tggtatatgg 780
aatcgccctg ccttctcatc ttcagcctct cgctgtacag gatccatcct cttagcactt 840
tttttcctct tagtaaacta ttgctgtgta acacagagaa atactcatag ctctgatcac 900
aaatctgagt cagatgcctc ttgaagaatg gacttctgag gagcaccgta ctcgccccag 960
cacagctctc acacaggagt ttgctgaaag tctcctgcct gtgtgcctcc cagtatgagc 1020
tctgtgaggg ccgtaacccc agggtataag taactcaact agaatatgtg caatacactt 1080
ctctgtgtcc tgatgttaaa cttaaaatga aata                            1114
```

<210> 15
<211> 1185
<212> DNA
<213> Homo sapiens

<400> 15

```
gtagacattt aagttttaaa tcctgtaaag agaagcagca gagatgcaac ataggctcag 60
actttaagca gctctaattt catggctttc ctgggcctcc accatattgc catagggcct 120
gtcacatgct tatgcatcct actggctgca gcagtgcatc ctgaagacag ctcagtatag 180
ccatgggcta cttttctagt catgggctgt ggaacctgga acatgtgacc agaataccac 240
tgaccctgtg gaccagcaag cctaagttaa aggcagccat catgaggatg aaaaggagac 300
cctcaccagc cagaagggga tgtatggggt aagcagaacc ctattagaga gaaaagccaa 360
caggaatgac tgggaaaaca gcaggagggg acaaaaattg gtgtgccttg gtgacacaaa 420
aaactctggg atgacggagt tgatgacaat cgaggccaaa atcatcttca acatacccta 480
gttctaaagt atgcactatt ctccccctca gcacagccct ttgggccttt tacacaatgg 540
tgacttctaa tttggtccct actcagattt tagaactccc aagtaatgtt cccaaatgga 600
tcatattcac gaggagtagt accagtgaat taactgggtc tcaaaataca attgcacatt 660
tcagacactg ctttaaagtt gtgtcctctc attttctcat tcctgttcag ttataatagg 720
cagtcttcct cttattcacc aggggattcc atccctctga ttcaggacca tggtatatgg 780
aatcgccctg ccttctcatc ttcagcctct cgctgtacag gatccatcct cttagcactt 840
tttttcctct tagtaaacta ttgctgtgta acacagagaa atactcatag ctctgatcac 900
aaatctgagt cagatgcctc ttgaagaatg gacttctgag gagcaccgta ctcgccccag 960
cacagctctc acacaggagt ttgctgaaag tctcctgcct gtgtgcctcc cagtatgagc 1020
```

tctgtgaggg ccgtaacccc agggtataag taactcaact agaatatgtg caatacactt 1080
ctctgtgtcc tgatgttaaa cttaaaatga aatagttttc agaataaaaa taaatttgta 1140
tatattaaaa aaacaaaaaa aaaaaagatc tttaattaag cggac 1185

<210> 16
<211> 413
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (138)
<223> a, c, g or t

<220>
<221> unsure
<222> (145)
<223> a, c, g or t

<220>
<221> unsure
<222> (151)
<223> a, c, g or t

<400> 16
acaaaatgtt tttatgtgtt taggtctgac ggtatgtctt cttattgttg acttttaaat 60
ttattaaatt gtagtttagg agaatgagct aatataatct ctaatagctt agaataattg 120
aagtaattta ttattttnat ttatnattat naccattatt attaacaact ttgtgaacaa 180
aagcatttat ttaaatgttt tatggttatt tgtattctct atttgtagga ttcaaagttt 240
tatattcaac ttattgagat tattatagaa atgaggtata tttgagaggg tcaaatattc 300
taaaagaata tttcagacac cacctcctcc agcaacaaaa atagctacaa accaatataa 360
attcatgtta gttattttca tacatattga aaactttact gtacattaaa agg 413

<210> 17
<211> 517
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (138)
<223> a, c, g or t

<220>
<221> unsure
<222> (145)
<223> a, c, g or t

<220>
<221> unsure
<222> (151)
<223> a, c, g or t

<400> 17
acaaaatgtt tttatgtgtt taggtctgac ggtatgtctt cttattgttg acttttaaat 60
ttattaaatt gtagtttagg agaatgagct aatataatct ctaatagctt agaataattg 120
aagtaattta ttattttnat ttatnattat naccattatt attaacaact ttgtgaacaa 180

aagcatttat ttaaatgttt tatggttatt tgtattctct atttgtagga ttcaaagttt 240
tatattcaac ttattgagat tattatagaa atgaggtata tttgagaggg tcaaatattc 300
taaaagaata tttcagacac cacctcctcc agcaacaaaa atagctacaa accaatataa 360
atttcatgtt agttattttc atacatattg aaaactttac tgtacattaa aaggttgtac 420
agacagccct ccatatctgt gggtttcaca tccatgtatt caaccaagca tggatcaaaa 480
atattcagta aaaaaaaagt cgacgcggcc gcgaatt 517

<210> 18
<211> 502
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (292)
<223> a, c, g or t

<400> 18
cacacagaat gcacaaaggc agggctgctc ttttattttg ggtggcgtgg gtggaagtag 60
gcaatggagg cacagttgtc aaagccactg gtctaatcta actattcttc tctcttccac 120
tttacagagg gtgtaacttg ggccaagaaa tcgtgatgtg agtaggtcac acagtgttga 180
gggccaaaac cctcgactcc gttgggagtt aaaggggagg aagtggacat ttcctttgaa 240
cctaataata agcaggtgag gttttgtatt ctcatttttc acagtgcaag anaattttgt 300
ttagaaattc tagttaagtc atttacccaa agccctaaag ccagtaaaca gcaagtggga 360
atcaaaccca agtcccttga catgccaaac cagcttcttt ccactagacc accccttccc 420
ccattactgt aaactccatc tctccgctgc tgtcagctcc ttcttgatcc gtacagaaat 480
aaaagtttgc caccaaaaaa aa 502

<210> 19
<211> 2961
<212> DNA
<213> Homo sapiens

<400> 19
cctgctgctc ctgggagggg ctcaggccat ttgattctgc ctggtgggtt cagggctgcc 60
tggctagcca ccaagacacg ggagagcaaa gcccggccaa acccaggggc agtggtgtcc 120
agcaggcatc aggacatgcc gtggcccaca catgcccttc tccacaggtg ccatcatctc 180
ctttcttctc ctgttagtga tggcttcatt ggcctcaggt gacaccgcct tctagcccag 240
gccactggat ccagccagag cgtctgtgca tcgccccagg gaggtgtggg gaggctcggc 300
cccacctggc aagtttatag gctgcttgag gacataagat gggtacagaa gaggcttcag 360
ggctcagcag atgttgtgct gggagtgagc ccaggcaagg tggcccatgg acagaagatg 420
gaaggtcagt cgtggcctcc ttccagaagg aaagagattt tagcaggaag agacaggcaa 480
ggaaaggaat tggtttttca ggcagagcgg tacttttta aattgacaaa ttttaagcat 540
acagaaaagt agagaggacg aacctccaaa catcagccct tagattccag tacatcctgc 600
cttggtagct gacgtcacga acgctgtgac ctgtccctct gcactccatc atagcatctc 660
taaaacggag gacatttcct acctcagcac aacttcatta tcacaaaaga gaacattttg 720
agcaaaagtt tggaagcagg caaaagtatg gtttgttctg gaatggcatg aatgaatgag 780
tcatgggctg gacccggccc cgcagaaaga gcagaggagg ccgtctctgg tgtaggagtt 840
gaagccaaga ctcagcatgc aggacaagga gcccagcctg gtggcatggg ctgtggcttc 900
tccagtgggc ccataggaat ggccctgggg ctgggacttg tggggacggc ggccacaaga 960
ggaggatctt cagcatggcc tgacagcacc tgtaatgtag gcaggcaatg ggctcctcca 1020
ggtggcagga acaccgtgag gagcatgcaa cgtgctggag accatggggc atgtgacctc 1080
cgggcacacc ctggacagac ctgggtgaga ggtgggctgg gacgccagga cagtgaaggg 1140
cttcaaggtg tgtttgtcct ttgtccttat acaggtgacc tgcatgggcg ggtgaggtcc 1200
atcagaatgc tctagaacag gacaactgag gtcccggctg ccccgctcaa cacgccatga 1260
ttcctagact gccctgagag gtcaggtttt aaatagggtg tggcccggca aggaccctga 1320
gggatgtgag gccagacagc tccttggtgg ccgcctgcac gggaagctga gtggcccttt 1380

ggaacaggaa gtgccttgtc caagttaacc cgggcccggg cccaccttgg actccctctg 1440
gagagaacgg ctggcctgtt cgcctccttt ctccttcctt ccaaaataac atgtccaggt 1500
cccttggtaa gaatgctgca gctggggtca gtctccaaac tggagtttgg tggctaaaat 1560
gcccgccttt ctgcccccaa tttcacacat ggatgatggt cctggttcag cagggaagat 1620
tctcttggac ccgtgtctgc tggttgggaa gtggagaaag ccgccctgtt ctgcctccat 1680
tttgaaacgt cttttctgtt cggaggacta aaggaagcca gagctggttc agatgttcat 1740
gagcacctgt gccagcccct gcccgtcaca cttgaggaca ccaaggctgt ggtctgtgca 1800
gtgaatcggg gcagtggggg gacttcatgg ttagaaacac aagctggagg ccggatgcgg 1860
tggctcatgc ctgtaatccc tgcactttgg gatgcctgag gtcaggagtt caagaccagc 1920
ctggccagca tggtgaaacc ctgtctctac taaaaataca aaataaaaat taactgggtg 1980
tgtggcaggc acctgtaatc ccagctactt gggaggctga ggcaggaaaa tcacttgaac 2040
ccgggaggtg gaggttgcag tgagctgaga tcgcaccatt gcactccagc ctgggcaaca 2100
agagtgaacc tccgtctcaa aaacaaaaca aaacaaaaag tagattctaa ttcaggaggt 2160
tcggagtgtg tgcatttcta atgcgctcca aggtgctgtt gctgctgctt agaaccatca 2220
tttcagtagt aagggtctaa aacaacacga ctcacaagac ctcaacccac cagcccaggt 2280
ggaaaccaat tcacaaacgt ctcagtgaaa tgattctgtt gatgctggtg cgcacttcta 2340
catttatgtg gggaaggagg tcaggacaga agttgaggga cttgggaccc caaagcagat 2400
aaggttgtgg ggtgcagatc ttgctctagg atcctgaagc gagtctctgg aggaggtcag 2460
gcaaaatggt ggcaaacttt tatttctgta cggatcaaga aggagctgac agcagcggag 2520
agatggagtt tacagtaatg ggggaagggg tggtctagtg gaaagaagct ggtttggcat 2580
gtcaagggac ttgggtttga ttcccacttg ctgtttactg gctttagggc tttgggtaaa 2640
tgacttaact agaatttcta aacaaaattt tcttgcactg tgaaaaatga gaatacaaaa 2700
cctcacctgc ttattattag gttcaaagga aatgtccact tcctcccctt taactcccaa 2760
cggagtcgag ggttttggcc ctcaacactg tgtgacctac tcacatcacg atttcttggc 2820
ccaagttaca ccctctgtaa agtggaagag agaagaatag ttagattaga ccagtggctt 2880
tgacaactgt gcctccattg cctacttcca cccacgccac ccaaaataaa agagcagccc 2940
tgcctttgtg cattctgtgt g 2961

<210> 20
<211> 653
<212> DNA
<213> Homo sapiens

<400> 20
gtagaacaga aagcctccct gctcactggc ctttcccctt ccctcactgg ctgcccatgg 60
agccaggatc agcatgggaa caatcaagga ggcaggggct tatcagtact atggacccct 120
acactggctc tgcctggtgg ttcttctctt cgcataccaa agacagaaat taagcctcca 180
agagtggtaa ctgacctcgg tcacacttgg tgggtgtggg aaaggattca aatgtaggtc 240
tgttctcttc ttcatctatc atggtccctg tcctggaggc aagtcgtctg gggctcagaa 300
aacacccctg ttgccactga ttggaattcc aagggtctgg gtgaagtggg gatgggcctc 360
cagcttgcct ccagcctgaa aaaatagtag agggtgttga ggctgggaag ggaggtgggg 420
ctcatgttgt acagggcctg agccagggag cttgggcttc attctgagta ctgtggcagc 480
cctgggaagg tttgtagcac agagggatat tgtcattttt ggaaagatcc ctttggctgc 540
ttaggtagag aagggcttca agagggcagg aagggacagg actacagaag ggctgcttc 600
agagtccaga ttaaggaagg agaggcctgg gcagtcagaa aggaagagaa gcc 653

<210> 21
<211> 765
<212> DNA
<213> Homo sapiens

<400> 21
gaacagaaag cctccctgct cactggcctt tccccttccc tcactggctg cccatggagc 60
caggatcagc atgggaacaa tcaaggaggc aggggcttat cagtactatg gacccctaca 120
ctggctctgc ctggtggttc ttctcttcgc ataccaaaga cagaaattaa gcctccaaga 180
gtggtaactg acctcggtca cacttggtgg gtgtgggaaa ggattcaaat gtaggtctgt 240
tctcttcttc atctatcatg gtccctgtcc tggaggcaag tcgtctgggg ctcagaaaac 300

acccctgttg ccactgattg gaattccaag ggtctgggtg aagtggggat gggcctccag 360
cttgcctcca gcctgaaaaa atagtagagg gtgttgaggc tgggaaggga ggtggggctc 420
atgttgtaca gggcctgagc cagggagctt gggcttcatt ctgagtactg tggcagccct 480
gggaaggttt gtagcacaga gggatattgt catttttgga aagatccctt tggctgctta 540
ggtagagaag ggcttcaaga gggcaggaag ggacaggact acagaagggg ctgcttcaga 600
gtccagatta aggaaggaga ggcctgggca gtcagaaagg aagagaagcc ggatgtggtg 660
gcttgtatct tcaatcccag cacttttgag aggccaaggt ggggaggatc actttgagcc 720
caggagtttc aaaacccagt tttggcaaca tagttagact gcttg 765

<210> 22
<211> 148
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (9)
<223> a, c, g or t

<400> 22
ctcggaatnc ggctcgagga gaatttttgg agatccttag tctgctattc ttactgacat 60
tagccttatg ttgttgatag tagtatagaa tagtattcga aatgatgtgg acaattttgt 120
caaatgcttt taagatattg tacatctg 148

<210> 23
<211> 398
<212> DNA
<213> Homo sapiens

<400> 23
ctctggacca attattgatt tgcaagttcc tgtgaaatgt tttatcaccc agtgtgtttt 60
gtctgtgtac aggaagcact ggatacactg gctaggtgtg tgcttttaag taacattctt 120
ctctgattcc tttccttcaa agtgggagat actggtagca tctacctacc ctgcaaggat 180
ttaatgagtt ttaatgacat gagaaatgct tagatttcag ggatttgact aaacccaaat 240
acctgggcca tatttttagc caggagccag ggaatatttg cactaagcca ctggctgacc 300
caaaacttct ctctttaagt catttagcca aacagcaaat tagggagcca gcagctcatt 360
ttgggggtga tgttttgtat gagaagttgg tgaacctg 398

<210> 24
<211> 523
<212> DNA
<213> Homo sapiens

<400> 24
tacaacttta gtttttgtgc tacttaggag agaaaagcag atattgcctt attttgtgtg 60
ccctatccat ttaattagaa gctcaatgaa aatttttatc attatattat cacctctatg 120
tggaatattg ttaaatgttt tagaaagttt gaaattcatt tttaaatgtg aatcattatt 180
gtttgtgtgg ggggaggaat gtcaagttgg aattatgaat caggctttgc cttaccaagt 240
acttttatat taatagggat tattttgaat ttctgtaata cgtatgtatc ctctataata 300
catagagaat gcaaaaggaa aataatttga aagctatcct attttatctt gaaagcaatc 360
ttttaaaagc gcattataca tttatttgaa cgcttatttg gaattgtctc ttttccattc 420
tcttttcctt taaagaaatg gtaggggaaa aacatctggt ataccctata gctgtcttac 480
tgtagtcaag gtaatatact acgaaagaac acacgtaccc tgg 523

<210> 25

<211> 5982
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (5780)
<223> a, c, g or t

<220>
<221> unsure
<222> (5885)
<223> a, c, g or t

<400> 25
atgggcccgc gagccaggcc ggcgcttctc ctcctgatgc ttttgcagac cgcggtcctg 60
caggggcgct tgctgcgtga gtccgagggc tgcgggcgaa ctagggggcgc ggcgggggtg 120
gaaaaatcga aactagcttt ttctttgcgc ttgggagttt gctaactttg gaggacctgc 180
tcaaccctat ccgcaagccc ctctccctac tttctgcgtc cagaccccgt gagggagtgc 240
ctaccactga actgcagata ggggtccctc gccccaggac ctgccccctc ccccggctgt 300
cccggctctg cggagtgact tttggaaccg cccactccct tcccccaact agaatgcttt 360
taaataaatc tcgtagttcc tcacttgagc tgagctaagc ctggggctcc ttgaacctgg 420
aactcgggtt tatttccaat gtcagctgtg cagttttttc cccagtcatc tccaaacagg 480
aagttcttcc ctgagtgctt gccgagaagg ctgagcaaac ccacagcagg atccgcacgg 540
ggtttccacc tcagaacgaa tgcgttgggc ggtgggggcg cgaaagagtg gcgttgggga 600
tctgaattct tcaccattcc acccactttt ggtgagacct ggggtggagg tctctagggt 660
gggaggctcc tgagagaggc ctacctcggg cctttcccca ctcttggcaa ttgttctttt 720
gcctggaaaa ttaagtatat gttagttttg aacgtttgaa ctgaacaatt ctcttttcgg 780
ctaggcttta ttgatttgca atgtgctgtg taattaagag gcctctctac aaagtactga 840
taatgaacat gtaagcaatg cactcacttc taagttacat tcatatctga tcttatttga 900
ttttcactag gcatagggag gtaggagcta ataatacgtt tattttacta gaagttaact 960
ggaattcaga ttatataact cttttcaggt tacaaagaac ataaataatc tggttttctg 1020
atgttatttc aagtactaca gctgcttcta atcttagttg acagtgattt tgccctgtag 1080
tgtagcacag tgttctgtgg gtcacacgcc ggcctcagca cagcactttg agttttggta 1140
ctacgtgtat ccacatttta cacatgacaa gaatgaggca tggcacggcc tgcttcctgg 1200
caaatttatt caatggtaca ctgggctttg gtggcagagc tcatgtctcc acttcatagc 1260
tatgattctt aaacatcaca ctgcattaga ggttgaataa taaaatttca tgttgagcag 1320
aaatattcat tgtttacaag tgtaaatgag tcccagccat gtgttgcact gttcaagccc 1380
caagggagag agcagggaaa caagtcttta ccctttgata ttttgcattc tagtgggaga 1440
gatgacaata agcaaatgag cagaaagata tacaacatca ggaaatcatg ggtgttgtga 1500
gaagcagaga agtcagggca agtcactctg ggctgacac ttgagcagag acatgaagga 1560
aataagaatg atattgactg ggagcagtat ttcccaggca aactgagtgg gcctggcaag 1620
ttggattaaa aagcgggttt tctcagcact actcatgtgt gtgtgtgtgg gggggggggg 1680
cggcgtgggg gtgggaaggg ggactaccat ctgcatgtag gatgtctagc agtatcctgt 1740
cctccctact cactaggtgc taggagcact cccccagtct tgacaaccaa aaatgtctct 1800
aaactttgcc acatgtcacc tagtagacaa actcctggtt aagaagctcg ggttgaaaaa 1860
aataaacaag tagtgctggg gagtagaggc caagaagtag gtaatgggct cagaagagga 1920
gccacaaaca aggttgtgca ggcgcctgta ggctgtggtg tgaattctag ccaaggagta 1980
acagtgatct gtcacaggct tttaaaagat tgctctggct gctatgtgga aagcagaatg 2040
aagggagcaa cagtaaaagc agggagccca gccaggaagc tgttacacag tccaggcaag 2100
aggtagtgga gtgggctggg tgggaacaga aaagggagtg acaaaccatt gtctcctgaa 2160
tatattctga aggaagttgc tgaaggattc tatgttgtgt gagagaaaga gaagaattgg 2220
ctgggtgtag tagctcatgc caaggaggag gccaaggaga gcagattcct gagctcagga 2280
gttcaagacc agcctgggca acacagcaaa accccttctc tacaaaaaat acaaaaatta 2340
gctgggtgtg gtggcatgca cctgtgatcc tagctactcg ggaggctgag gtggagggta 2400
ttgcttgagc ccaggaagtt gaggctgcag tgagccatga ctgtgccact gtacttcagc 2460
ctaggtgaca gagcaagacc ctgtctcccc tgaccccctg aaaaagagaa gagttaaagt 2520
tgactttgtt ctttatttta attttattgg cctgagcagt ggggtaattg gcaatgccat 2580
ttctgagatg gtgaaggcag aggaaagagc agtttggggt aaatcaagga tctgcatttg 2640

```
ggacatgtta agtttgagat tccagtcagg cttccaagtg gtgaggccac ataggcagtt 2700
cagtgtaaga attcaggacc aaggctgggc acggtggctc acttctgtaa tcccagcact 2760
ttggtggctg aggcaggtag atcatttgag gtcaggagtt tgagacaagc ttggccaaca 2820
tggtgaaacc ccatgtctac taaaaataca aaaattagcc tggtgtggtg gcgcacgcct 2880
atagtcccag gttttcagga ggcttaggta ggagaatccc ttgaacccag gaggtgcagg 2940
ttgcagtgag ctgagattgt gccactgcac tccagcctgg gtgatagagt gagactctgt 3000
ctcaaaaaaa aaaaaaaaaa aaaaaaaaaa aactgaagga attattcctc aggatttggg 3060
tctaatttgc cctgagcacc aactcctgag ttcaactacc atggctagac acaccttaac 3120
attttctaga atccaccagc tttagtggag tctgtctaat catgagtatt ggaataggat 3180
ctgggggcag tgagggggtg gcagccacgt gtggcagaga aaagcacaca aggaaagagc 3240
acccaggact gtcatatgga agaaagacag gactgcaact caccccttcac aaaatgagga 3300
ccagacacag ctgatggtat gagttgatgc aggtgtgtgg agcctcaaca tcctgctccc 3360
ctcctactac acatggttaa ggcctgttgc tctgtctcca ggttcacact ctctgcacta 3420
cctcttcatg ggtgcctcag agcaggacct tggtctttcc ttgtttgaag ctttgggcta 3480
cgtggatgac cagctgttcg tgttctatga tcatgagagt cgccgtgtgg agccccgaac 3540
tccatgggtt tccagtagaa tttcaagcca gatgtggctg cagctgagtc agagtctgaa 3600
agggtgggat cacatgttca ctgttgactt ctggactatt atggaaaatc acaaccacag 3660
caagggtatg tggagagggg gcctcacctt cctgaggttg tcagagcttt tcatcttttc 3720
atgcatcttg aaggaaacag ctggaagtct gaggtcttgt gggagcaggg aagagggaag 3780
gaatttgctt cctgagatca tttggtcctt ggggatggtg gaaataggga cctattcctt 3840
tggttgcagt taacaaggct ggggattttt ccagagtccc acaccctgca ggtcatcctg 3900
ggctgtgaaa tgcaagaaga caacagtacc gagggctact ggaagtacgg gtatgatggg 3960
caggaccacc ttgaattctg ccctgacaca ctggattgga gagcagcaga acccaggggcc 4020
tggcccacca agctggagtg ggaaaggcac aagattcggg ccaggcagaa cagggcctac 4080
ctggagaggg actgccctgc acagctgcag cagttgctgg agctggggag aggtgttttg 4140
gaccaacaag gtatggtgga aacacacttc tgcccctata ctctagtggc agagtggagg 4200
aggttgcagg gcacggaatc cctggttgga gtttcagagg tggctgaggc tgtgtgcctc 4260
tccaaattct gggaagggac tttctcaatc ctagagtctc taccttataa ttgagatgta 4320
tgagacagcc acaagtcatg ggtttaattt cttttctcca tgcatatggc tcaaagggaa 4380
gtgtctatgg cccttgcttt ttatttaacc aataatcttt tgtatattta tacctgttaa 4440
aaattcagaa atgtcaaggc cgggcacggt ggctcacccc tgtaatccca gcactttggg 4500
aggccgaggc cagcagatca cctgaggtca ggagtttgag accagcctga ccaacatggt 4560
gaaacccgtc tctaaaaaaa tacaaaaatt agctggtcac agtcatgcgc acctgtaatc 4620
ccagctactc gggaggctga ggcaggagaa ccgcttgaac ctgggaggta gaggctgcag 4680
gttagaagta cagaattcct gaagcacctt agggggtaag tcagtaggga cttaggtaag 4740
taacgtgtgg aaaagatagt ggtccattgt agagagagtg cactagaata caacttcggg 4800
tcagggaagc agcaattata ggtccaccta cagggaatga tgcagtctcc tgccttgggt 4860
aacatgttag tggcaaagct ttacagggtc caagcagggg acccacttca agagagtgac 4920
gattgcaatt aatgacacat aaaggaattg ttttttcctt atctaggtga ataggggatc 4980
ttgagtaagt gtaagtaact gatgacagcc ctggcttttg tctaacagta atataacaga 5040
gtaatagcta ctactaactg agttctccta tgtgtcaagc tctgtgcaag acactttaca 5100
atgtcatttg acttaattct cccctctcag agtcagaatg atcccaaatt tcataaagga 5160
ggaaaatgag ttcacaagaa ttacattcct taaggtaaca ctggtaagtg gctcagttgg 5220
gattcaaacc cagggtacgt gtgttctttc gtagtatatt accttgacta cagtaagaca 5280
gctatagggt ataccagatg tttttcccct accatttctt taaaggaaaa gagaatggaa 5340
aagagacaat tccaaataac tttcaaataa atgtataatg cgcttttaaa agattgcttt 5400
caagataaaa taggatagct ttcaaattat tttcctttg cattctctat gtattataga 5460
ggatacatac gtattacaga aattcaaaat aatccctatt aatataaaag tacttggtaa 5520
ggcaaagcct gattcataat tccaacttga cattcctccc cccacacaaa caataatgat 5580
tcacatttaa aaatgaattt caaactttct aaaacattta acaatattcc acatagaggt 5640
gataatataa tgataaaaat tttcattgag cttctaatta aatggatagg gcacacaaaa 5700
taaggcaata tctgcttttc tctcctaagt agcacaaaaa ctaaagttgt atagtgtact 5760
acttttggaa gagacatttn taactagtaa tgataatttg tcttaattca taaacacttc 5820
aaatcacata actgaataca ttttcaacca ggaggatgca acattacccc aaaataccga 5880
gtcanagaaa ttattattgt tgagacaacc aggtaccaaa ctcttaattc ccattgggtt 5940
tctgggcctt tttacctgta ctttaacact aattcccgtg aa                    5982
```

<210> 26

<211> 820
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (502)..(565)
<223> a, c, g or t

<400> 26
caaacaatat gtgtgagatt taatgatttt tgtcctattt tttttctatt tattatttaa 60
tactttcagg ggtatcctag taaaagcgac tttccatgtg cctgcctgtg cctttctgtg 120
ccagggtctt aagaaatgtg tgtgggtttc ctctgtggct ctgacactcc atcaaacacc 180
aggacttgtg cacgcaggtg agaagcatac ccaggtgcct cagagatatt gaagagggtc 240
tggaatgtgg gagagaggca aatggctttc caataaaagt aggctaagac ataattaggg 300
gcttaggtgc tgctttaaaa aataagtagc agtgtattcc agactcctct aagaaaagag 360
aattgctcat tgtggaacag gcatggcaat gcagtgcctt tgcctggaac accttgctgc 420
tgccatcact aagacccatt ctgggacaag aggaggctta ccccttattg agtatctgcc 480
atgagctttg tgtggtcctt gnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 540
nnnnnnnnnn nnnnnnnnnn nnnnnggcca cccatttcct gacagagatt ggcctacaaa 600
ggtcagatgg cttccagggc acactgagcg gcccttgtgt gtcggaatgt tcccttctgt 660
ttgtcccttc caggctggac actttgggag cagaagtcaa agacaccttt atcattgtac 720
cctcagcacc tggtgtagtg cctgggattt agtagttctg aggagcgtgt gttgaatgaa 780
tggaggttaa gtaacttaga cattagatag taggactctg 820

<210> 27
<211> 839
<212> DNA
<213> Homo sapiens

<400> 27
gaaatagagc acataaaacc aaacaatatg tgtgagattt aatgattttt gtcctatttt 60
ttttctattt attatttaat actttcaggg gtatcctagt aaaagcgact ttccatgtgc 120
ctgcctgtgc ctttctgtgc cagggtctta agaaatgtgt gtgggtttcc tctgtggctc 180
tgacactcca tcaaacacca gggcttgtgc acgcaggtga gaagcatacc caggtgcctc 240
agagatattg aagagggtct ggaatgtggg agagaggcaa atggctttcc aataaaagta 300
ggctaagaca taattagggg cttaggtgct gctttaaaaa ataagtagca gtgtattcca 360
gactcctcta agaaaagaga attgctcatt gtggaacagg catggcaatg cagtgccttt 420
gcctggaaca ccttgctgct gccatcacta agacccattc tgggacaaga ggaggcttac 480
cccttattga gtatctgcca tgagctttgt gtggtccttg cttgctaggt acttgagcaa 540
gcattatctt ttttaccttt taaaacacca ctgaggtgta ggtatggcac ccatttcctg 600
acagagattg gcctacaaag gtcagatggc ttccagggca cactgagcgg cccttgtgtg 660
tcggaatgtt cccttctgtt tgtcccttcc aggctggaca ctttgggagc agaagtcaaa 720
gacaccttta tcattgtacc ctcagcacct ggtgtagtgc ctgggattta gtagttctga 780
ggagcgtgtg ttgaatgaat ggaggttaag taacttagac attagatagt aggactctg 839

<210> 28
<211> 191
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (175)
<223> a, c, g or t

<400> 28

gacacttcat gttccagctt ctcgcttcaa gctcctgaag gtcagtgctt aagagcaggg 60
tcaaggagtg cagaacccat tggtgagatt tacaggaccc cccacccaac cgttagcatg 120
ccgtctgtag catgggaggg gcctggagca tagcagggcc tctcacgggc tttgntttca 180
ggttgacatt t 191

<210> 29
<211> 998
<212> DNA
<213> Homo sapiens

<400> 29
ggccgcatat tttttttttt tttttttttt ttttaaaggt agaaaataac acaacctttt 60
attttttatt ttcatttgtt ttgagatgga gtttcactct tgtcacccag gctggagtgc 120
aatggtgtga tctcggctca ctgcaacctc cacctcccag gctcaagtga ttctcctgcc 180
ttggcctccc aagtagctgg gattacaggc atgcactaca acgcccagct aatttttgca 240
tttttagtag agatagggtt tcaccatgtt ggccaggctg gtcctcaact cctgacttca 300
ggtgatccat ccatcttggc ctcccaaagt gctgggatta caggcgtgac agctgtgccg 360
ggcccacctt ttaaatgtca acctgaaacc aaagcccgtg agaggccctg ctatgctcca 420
ggcccctccc atgctacaga cggcatgcta acggttgggt ggggggtcct gtaaatctca 480
ccaatgggtt ctgcactcct tgaccctgct cttaagcact gaccttcagg agcttgaagc 540
gagaagctgg aacaatgaag tgtctattct gcttcttctt gcaaatgctg caactacaga 600
aagacagagc aaattccaga ttgtgagcag ccacctgcat cctctatgcc tgagcggccc 660
agccatgaga gccagccgac cccacagatg atgccccttt cagcaccatc cagggccgag 720
gagctggggc aaaggcctgg atagcagtgc ctctggtttg caggtacagc agagcccagg 780
ggggtcccaa gtcagcagtc gaggttctgc aatgctcaga acacaggacc aacagacagg 840
tctgtactgc ccaccccctca gttctttaca gtgaagagaa gcgctggact tcagagacac 900
ttaggaaaca aatttcagac actgctacaa cctgatgtct ctgagacatc cacaccaaaa 960
atggacaagg aagtctaggt ttccctcttt cttcatca 998

<210> 30
<211> 282
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (5)
<223> a, c, g or t

<220>
<221> unsure
<222> (17)
<223> a, c, g or t

<220>
<221> unsure
<222> (29)
<223> a, c, g or t

<220>
<221> unsure
<222> (110)
<223> a, c, g or t

<220>
<221> unsure
<222> (128)..(217)

<223> a, c, g or t

<400> 30
accangggtc cgaccanggg tacgaccang ggtccggctt gatctcactt atatatggaa 60
cttaaaaact acacatagaa acagggtaga atggtagtta tccaggctcn ggaggaagag 120
aaaacaannn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 180
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnngta tatggaaaat ttgctgagag 240
actagatttt aggtattcta cctcaattaa aaaggtaatt gt 282

<210> 31
<211> 1225
<212> DNA
<213> Homo sapiens

<400> 31
ggggaaagac caaccagttg gggctttctc ccagggctcc cgggggctgt gtctgagtgt 60
ctgtgttggc tgtttggttt gctttggttt tgtttctgga ggttgcttgc aggtttttgg 120
aggaagtgac tagtttggtt aagagctggg aactgagtca ggtaagccgt gtcatgttgt 180
aactccacca gaaaatggag gagagcgggt ttccaggaga caaagctgag atgagaagtg 240
tttataaaat tatggatgtc tccattttga agctctgttg gtgatggctg gaggaggagg 300
aggcttgctt gcctactcct tctctctttc cagagggaaa ccttgtggtg gttcctcact 360
gtctattcat tatgcaagga aatgagggct tttaagggtt cctcagattt ttctccacca 420
aagagtgctt tcacaagtta ttgaggcgtt tgtttccatt ttaaagtaaa cttttggaat 480
tttttttctc cttttgagtg gacctgaagg gttttgacct ccttcaggaa aggcaaggca 540
aaaacttaaa acagttcact gaggtctcac acaactttaa gctgctccag gtctcctgaa 600
agtcaccagg aaatgtgatt tcctccttgt gaagatggtg atggccctaa gctgagattt 660
ttttgagttc tagggtttgg ttatcatcat gttttgatgc attgcaagac tttattgtct 720
gatttgagtt gatttctgca aatataaaaa ataataagaa taatcctgca agatctcaga 780
ggaactctaa gactggctaa caccagtttc tccaggttct ccatttctct tcaggtcgtt 840
ccatttgtat gttaggcctt ctttcagttt ctttgtttcc cctttccctt tccgtcggct 900
aatttttttct gtgttctgaa gtactcttaa gtcttcagaa atatcagtat gtcttcttaa 960
caatgtcgct atggaaacaa attttaaaaa catgatgtca gttgagaaaa ccttatgtcc 1020
aggtatcttc acctttttaa ttgggaggaa tttattaatc atgtaggaag acattttgtg 1080
aggataattt gaaaaaagga cccagtgcta ccctagtcca cacacattga tgggagctct 1140
tcacatatta gttttagaga atgtacataa ttgacccaag caaagaacta aatcccgaaa 1200
tgcttcagga atttttaaaa gccaa 1225

<210> 32
<211> 844
<212> DNA
<213> Homo sapiens

<400> 32
tctgccatgg tgacaattat tttaaattta gatgagatca tcataacatg acacataaaa 60
attgtttcat catgcatatc aagtttgata tgtagcttaa atttattttg cacacactag 120
aatttgtcct ggttttctag tacctcaagg cagatatgca aaggtgttta ggagacatac 180
tctcagacaa accattatta ttttaaagga tagaacaaaa caatcgctag ttaaggaaga 240
tgttttgtaa taattaaact tgtaattatt tgacttgaaa tatttaatca ttttttttggg 300
aaagaatgga tagattttgt taatgttagc actcttaaaa ttaagcagtg gcttttttcc 360
ccgtgtctcc catattctcc ttgtgtttga aacataaaac aaacactaaa cctaagcaaa 420
agttgctggg tttgttttca taattgaggt gagtttttcc ctcaactatt acaataaaag 480
aaaacttttt atgattttaa tgataatgtt ttgtggtggg ttaaagacct cctaacaaca 540
gggggttttt atacaacaac aagaagtttt taaataattg agtttttaaa gtggaaagca 600
gcagtaaatt aaactagaag gatatatttt atacctagaa ataaataaag ctcaacttgt 660
tttgtaagcc tgttttaaaa atatttaatc atttaatttg tgcaagtata gagttctcct 720
atggcaaaac tataccatca tcttctccaa ttgtgcatgg cagctgtact aagttctgca 780
aaaacaagac atatggatgt gtttcatacc ttctcagaat tggtatatca agacacattt 840

aaat 844

<210> 33
<211> 2483
<212> DNA
<213> Homo sapiens

<400> 33
gatctagaac tagtcatcgc gagcggcctt tttttttttt ttttttaaga tggagtttcg 60
ctcttgttgc ccaggttgga gtgcaatggc gcgatcttgg cttactgcaa cctctgcctt 120
gcagttcaag caattctact gcctcagcct ctcaagtagc tgggactata gacattcacc 180
accacaccca gctaattttt tgtattttta gaaaattttg tatatttaga aaaggtttca 240
ccatgttggc caggctggtc ttgcactcct gacctcaggt gatccgccca cctcagcctc 300
ccaaagtgct gggattacag gcctgagcca ctgtgcccag ccctcaagta actcttaaac 360
ctactgaagt tagacaatca ataactgaaa tgacatcatc tttcttgaat gtttaaggaa 420
ataaagttcc ttcttctgac aaactttaag tgtgttcttg atttccttgc ctccctcttc 480
ctctgggagt tttcttccct agctgctcac tttcattatc aacgaaatat tcctcttcac 540
gcctttttac cttataccta caacatgctc agttctctct ctttacaaga aaatataagt 600
tttcaccaac ctatttatca aatttacatc cccctccctt tctacttcct tttgtaaaaa 660
aagagcattc aacctattgt ctgtctccat gccctcacat tatcagtgca agcacccgca 720
actgtggctc tccaccatgt gagctcaacc tatcatcaca actgtatctc ccctaacact 780
catttagatt aagccatttt tcacaagttt ctaaaattat ctcttccatt tctcagtata 840
accctttctt tcccttcaca gtttcttgaa ccaatctcac tagtccttca acgttcactt 900
ccaaggccac cccgaacaca tcttttcctc ttccctaaat aaattctact ggattctttc 960
tgtttttcac tggaaacttc tcatactcca ttggttcctt tctcatgaca tttattttac 1020
atccctagta ttctggtttc ttacattttt tttcctatct actaaacaat aacttctttg 1080
agaactggac cagtgtctct tatatttata tccctaataa tacttattaa acacgtagtc 1140
tattctcaac attgaattcc atcttatact caaagaataa tactttaaca tagccattgt 1200
tcatagtgta tatataatta agaacacatt ccatattttt cttgagatta tatagtgtta 1260
aatttttcaa aattatagga tatgatctaa agatatattt taaaactcaa acctgtaatt 1320
ttatcttcag ttatgctata gcatgtacat ttccattctc ttgtcgaagt ttctttcgtt 1380
cctcagcttc tccttcatat ttcctgacgt attgtcttct aagccttcag agaacaaggc 1440
attctaatgt tatcaaggtt ctattcatct atatgtttga ttggggtttt tatgagtaga 1500
gggggttcca cttcatgagt agtgatagac cagcaatcac tatacttgac actaaaccta 1560
aacctggcta taaaatatta ccaatttcta agggggtatt tatgttgact gtatataaat 1620
ccatttccag agggcttata tttaaatgtg tcttgatata ccaattctga gaaggtatga 1680
aacacatcca tatgtcttgt ttttgcagaa cttagtacag ctgccatgca caattggaga 1740
agatgatggt atagttttgc cataggagaa ctctatactt gcacaaatta aatgattaaa 1800
tatttttaaa acaggcttac aaaacaagtt gagctttatt tatttctagg tataaaatat 1860
atccttctag tttaatttac tgctgctttc cactttaaaa actcaattat ttaaaaactt 1920
cttgttgttg tataaaaacc ccctgttgtt aggaggtctt taacccacca caaaacatta 1980
tcattaaaat cataaaaagt tttcttttat tgtaatagtt gagggaaaaa ctcacctcaa 2040
ttatgaaaac aaacccagca acttttgctt aggtttagtg tttgttttat gtttcaaaca 2100
caaggagaat atgggagaca cggggaaaaa agccactgct taattttaag agtgctaaca 2160
ttaacaaaat ctatccattc tttcccaaaa aaatgattaa atatttcaag tcaaataatt 2220
acaagtttaa ttattacaaa acatcttcct taactagcga ttgttttgtt ctatccttta 2280
aaataataat ggtttgtctg agagtatgtc tcctaaacac ctttgcatat ctgccttgag 2340
gtactagaaa accaggacaa attctagtgt gtgcaaaata aatttaagct acatatcaaa 2400
cttgatatgc atgatgaaac aatttttatg tgtcatgtta tgatgatctc atctaaattt 2460
aaaataattg tcaccatggc aga 2483

<210> 34
<211> 591
<212> DNA
<213> Homo sapiens

<400> 34

aatccattta aagtcaacta aatgttattt gggtaatatg tccctcatgt taaatttgcc 60
caaatatata tctcacctct taaaattcgt ttagtttgaa attaaaatta gtattgtttt 120
tctgcatgta ctcctaggtt gggtaaagaa gggaacaagg gaatggggaa acgtagagat 180
tcctggacta acagagaaag acagcttgag aataaaagta tgcaaaagat aatctacaac 240
aaaataatgc acttaactct tgttactaaa caaataagct acccacattt cagcttatct 300
gtatttgttt catgatttgt cagctatcta gcaactatct tagtcactga ttcggaacga 360
cttagcagtg gttattgcat agaacaactc cttacacaga gatttgcaag ctttctgaac 420
tttcgtactt tcaaattgaa aatcaggaga aacattttca acggcttcat attcagacca 480
agattagtat attaacaact aataacaata ttaaaagtta gaacaattcc tttcctctat 540
ctttctcagg acaaactcga gcttattaga aaactaggga gtgatctggt g 591

<210> 35
<211> 306
<212> DNA
<213> Homo sapiens

<400> 35
cctagggaga atcagcaagg aaaagatcaa tgtaatcttg aataacttat cctgaaactt 60
ctccagagtt acccagagag tcaacagtca tgctgctttt tgtacttagt ctggtgtttc 120
agtaccagtt taacacataa aaagtgatca aggtgcaagg gacacagctt tgaaatagtc 180
agacctggat ctgaatctgt gattctgtca tctgcaataa gtttctaact tctccaagcc 240
ttagtttttt atctgtaaag gggagtatta actagagatg aggattaaat gaaaagtcac 300
ttactc 306

<210> 36
<211> 617
<212> DNA
<213> Homo sapiens

<400> 36
ccaagactga gttagatttt ctattatgta ctcccatggc aacagcattt tccacttaac 60
ttgttggaaa agggacaact gtcctctggg ggctctgttg ccaatatttg ttccactttc 120
tctttcattt tcactttctt ccttacactt gcaatccaga gtccagatgt aaaacagtgt 180
agggccataa gtgatgggac atctctaaca aaattcttgg aggctgctgc ctggaaactt 240
gtgtccttgg gatggtaccc ttacccctga ggtgctaggg atgggcccca gggtctttcc 300
ctgctttcta ctttcctaat ggctaagtga tgtcagagga caacatcttg atgtgtagag 360
gtacaagaat tcagggatgc aaggatgcct tcctgcaaga cagagatcat tctatctaaa 420
ccaatgtttt caggtttttt actaggagca catgcatgaa tgtgtatata tgtgtatagc 480
tatgcaaaaa catgaacaga tgtatgcatg tgtataatct aaaacacata aaggtacata 540
tactgacata ctgaaacaca tattaatata accaaaaata aaaatttcat gagacagtat 600
taatgtttac cacatgc 617

<210> 37
<211> 725
<212> DNA
<213> Homo sapiens

<400> 37
ccaagactga gttagatttt ctattatgta ctcccatggc aacagcattt tccacttaac 60
ttgttggaaa agggacaact gtcctctggg ggctctgttg ccaatatttg ttccactttc 120
tctttcattt tcactttctt ccttacactt gcaatccaga gtccagatgt aaaacagtgt 180
agggccataa gtgatgggac atctctaaca aaattcttgg aggctgctgc ctggaaactt 240
gtgtccttgg gatggtaccc ttacccctga ggtgctaggg atgggcccca gggtctttcc 300
ctgctttcta ctttcctaat ggctaagtga tgtcagagga caacatcttg atgtgtagag 360
gtacaagaat tcagggatgc aaggatgcct tcctgcaaga cagagatcat tctatctaaa 420
ccaattgttt tcaggttttt tactaggagc acatgcatga atgtgtatat atgtgtatag 480

```
ctatgcaaaa acatgaacag atgtatgcat gtgtataatc taaaacacat aaaggtacat 540
atactgacat actgaaacac atattaatat aaccaaaata aaaatttcat gagacagtat 600
taatgttaac cacatgctat atacttatat ttttctttca tttgcaaaag aatgctgtta 660
tgactgtcta aacctctggc ttgagaaaaa aaaaaaaaaa aaaaagatct ttaattaagc 720
gtgcc                                                              725


<210> 38
<211> 90
<212> DNA
<213> Homo sapiens

<400> 38
gtaaaatatc tgtctcactg gcaatttttt ttacattgaa tttgttgaca attttttac 60
attgaatatg ttaaaatttt tatatattgg                                 90


<210> 39
<211> 222
<212> DNA
<213> Homo sapiens

<400> 39
tgtagagatg ggatctctct ttgttgcccg ggctggtctg gaattctctg ggttcaggtg 60
atcctgctac gtcagccatg agccacggtg cccagcctgg caggcttggt ttctcttaat 120
gcctctcctt ggcttgcaag atggccacct tctggctgtg tcctctctct catggccttt 180
cctttgtggg cacacatcct tgttctctcc ttcttcttat aa                    222


<210> 40
<211> 257
<212> DNA
<213> Homo sapiens

<400> 40
gttttcccat tgactaacgc ttaagatata ttggagtcaa atgctcataa aatgctcatc 60
caatgcttat aaaatattag agttgaaatg gactctctgt tcatgcagat gatgagaccg 120
aaacagagag cttccaggag gatcaatgcc attcaatgag cttgctgctg tactcccctc 180
tacacaatat ggatatatcc catcccagcc cgagactggc catactagtt ctagtaactg 240
aggctttcct cctactt                                                257


<210> 41
<211> 263
<212> DNA
<213> Homo sapiens

<400> 41
gctcgaggtt ttcccattga ctaacgctta agatatattg gagtcaaatg ctcataaaat 60
gctcatcaat gcttataaaa tattagagtt gaaatggact ctctgttcat gcagatgatg 120
agaccgaaac agagagcttc caggaggatc aatgccattc aatgagcttg ctgctgtact 180
cccctctaca caatatggat atatcccatc ccagcccgag actggccata ctagttctag 240
taactgaggc tttcctccta ctt                                         263


<210> 42
<211> 533
<212> DNA
<213> Homo sapiens
```

<220>
<221> unsure
<222> (501)
<223> a, c, g or t

<220>
<221> unsure
<222> (514)
<223> a, c, g or t

<220>
<221> unsure
<222> (528)
<223> a, c, g or t

<400> 42
atacagagtc gtgttggatt ggcagtctta aatcactctt gttatttcca gtggacatta 60
aaaaaaaatc acagataagt acttaaaaca ctcaagattt gggatttaga tcatgattag 120
atacaataga aagatcctgg aatcccgaca tgaggacaaa aatggtactg aattcttttt 180
gaaaaataga ttactgaaaa gcgatctaat atagaacagt tgcttttact tagatgttca 240
atgcatattt gttgtataat aaccaagtta ttacagttca gataaagggt ccaaagtgtt 300
ttcgttatga tataatactt tctattgtaa actggactaa agaaacgttg tatgttcaag 360
gaagtgttga gcagccatgg tgttcctggg acatgctccc caggtgctga gagaggtgct 420
gcaggagtca cagacctgca ggcacgcact tgccagtgac tgggacgttg gctggtggtt 480
ctcttttggt gtgattagag ntatgtgagt tgtntcaata cttgagantg tcg 533

<210> 43
<211> 676
<212> DNA
<213> Homo sapiens

<400> 43
atacagagtc gtgttggatt ggcagtctta aatcactctt gttatttcca gtggacatta 60
aaaaaaaatc acagataagt acttaaaaca ctcaagattt gggatttaga tcatgattag 120
atacaataga aagatcctgg aatcccgaca tgaggacaaa aatggtactg aattcttttt 180
gaaaaataga ttactgaaaa gcgatctaat atagaacagt tgcttttact tagatgttca 240
atgcatattt gttgtataat aaccaagtta ttacagttca gataaagggt ccaaagtgtt 300
ttcgttatga tataatactt tctattgtaa actggactaa agaaacgttg tatgttcaag 360
gaagtgttga gcagccatgg tgttcctggg acaggctccc caggtgctga gagaggtgct 420
gcaggagtca cagacctgca ggcacgcact tgccagtgac tgggacgttg gctggtggtt 480
ctcttttggt gtgattagag ctatgctgta caggaagcat ggctggggag gcctcgggaa 540
acttacaatc atggtgaaag gcaaagggga agcaggtttg tcccataatt cttcgggcct 600
ctctcaagcc ttcgagtgga tgctgtttca tatttcatcc agcctgggag ttggagacct 660
gagctgcatt acctaa 676

<210> 44
<211> 251
<212> DNA
<213> Homo sapiens

<400> 44
caggcctgct cagcaagatt ttcatgggat tagtgaattg gtggttgcca aatgccataa 60
taatgcacca tgcagtagac ttgctgtaaa gcacagtttc atcataacaa taactgtaaa 120
taatgctact gaacagctac agagcactcc tctgaactca ctggaatggg ctatatccca 180
tggcaagatg agtaagcctc aagcgcaaaa atctcaccct tgtttccctt ttttttggc 240
agaaatcccg a 251

<210> 45
<211> 606
<212> DNA
<213> Homo sapiens

<400> 45
agcgcccatg ggattaggga gagcatggcc ttcagaggct ggagctgtag tcctaactgc 60
acagctggtc cagcagggcg taacgcatct acctagagag taaaatgaca acagttgttc 120
cctaagctca gcacttgcaa agaaatcttt tgggaagatc tcttcaaatg tctagaactc 180
tgcgcaaaca ataggtagga caagtgtgaa cctacccaac ctctgttgac aaatacagct 240
gcacacccct cagcgaggcc tgctgtgaaa tgccaccttg gtgaaaatga gaataaaggg 300
tgagtgagcc agctgctttt ggatgaccaa attaattctt agcctcccat taagacaggc 360
ctgctcagca agattttcat gggattagtg aattggtggt tgccaaatgc cataataatg 420
caccatgcag tagacttgct gtaaagcaca gtttcatcat aacaataact gtaaataatg 480
ctactgaaca agctacagag cactcctctg aactcactgg aatgggctat atcccatgca 540
agatgagtaa gcctcaagcg caaaaatctc acccttgttt ccctttttttt ttggcagaaa 600
tcccga 606

<210> 46
<211> 455
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (14)
<223> a, c, g or t

<220>
<221> unsure
<222> (16)
<223> a, c, g or t

<220>
<221> unsure
<222> (18)
<223> a, c, g or t

<400> 46
agaaaaggag agananangg agggagggag aaagaaagaa ggtaaattga gacagcacta 60
agtgctatgg agagaattaa aggtgagcca tgtaatgtgg ctgaagaatt aggttaaatt 120
gggtgttccc atggcctaag gggtgacat ttaagttagg gttaacatgg agaggtgagc 180
aggaagagag tttcaaatca tgtgagagct agtcccaatg ccgtaaggag gaaatgggat 240
tggtgtgttt gaggaactga ggaagggcag ctgggagtat ggtacatgaa aaagaccgtg 300
acaaagaata agattgggca catagatggc agttccatct tctcacgttg tatgccaaag 360
taagaagttc ctaatttact gatagcaatg tgaacccaat gagaaacttt taaaagaaga 420
atgaagccat ctggttaagt atttaaaagt tcatt 455

<210> 47
<211> 367
<212> DNA
<213> Homo sapiens

<400> 47
catctggtgg cttctttcgt gcaacacatt cttaccaaat gcccgctgca caccacgcat 60

gcactgtcct gagcgctagg aacagcagca accaggaccg gcagggctca gagcctcccc 120
tgtgtgccaa gaatacagac agcccaggca gagggcattc ggtgctccag acacaaagtg 180
aaggcccagc ttcaaatgtg gctggatcca ggcacacatc ctgaggttct gctggtctgg 240
actgctaacc cactcacgag gatccattct caagcagccc cagcctgtct ccccacctgg 300
gcacgtcatg gctggggtct cttgatgggc aaggctccat tgatcgagtc ccctttggac 360
tggggca 367

<210> 48
<211> 249
<212> DNA
<213> Homo sapiens

<400> 48
gttttattgt tttcatgcat tcttaggtct tctactgcat gcttcctaaa aacagggtca 60
tcctgtctca gacctttcac tgaactcaaa ctgtgcttct cttgccagtc tcccgatgaa 120
gggcctgcca gggaaataaa cttggttgag acaaaattct tgtaaataag ctcatagagg 180
ggacagactc ctgctccatt cctcccaccc ctcacaaggt cttccaaatt agcggaaaac 240
agtctaaat 249

<210> 49
<211> 436
<212> DNA
<213> Homo sapiens

<400> 49
aaaatgacct tcttttgctg ttttcctgac cacaaggtgg attctgccca tcagggcctc 60
cctttacctt ggcacacaag ggcctccatc ctggggcacc ctcccacctg ccccacccct 120
ccaacctaac tacttgccat ttcccaaata catcgtgcag tgcctggtgt ctatcctgtg 180
gttcatgctg cttcctctgc ctgcaatatc ctttccctgc aatagcctct tccacccagc 240
agacgcctcc tcactctctt ggctcagctc taagtcatat cccctaggga agcttaccag 300
gatgctgcag tcagatggag tgtctcctcc tgggccccca cagaccctgt acttccttct 360
gtgacagcaa caatcccaca gcagtggaac tgcttacctg cctgagctcc tggagggcaa 420
ggatgccatg ttgctt 436

<210> 50
<211> 853
<212> DNA
<213> Homo sapiens

<400> 50
cctggataat aaatgcctaa ctctcaccaa ctaccttgcc atcatgggct tcttcgatca 60
gaaatctagt aagagagtat ggtgggggct gcgagaccca tcatcactac caaaaaacat 120
gaagtctttt cactttcaat atgtgaagac ctaaattatg taattggatg aatgatattt 180
gtagagtcaa agagaatgtg agttctccgt cttatgggta tagttattat gtaataatca 240
agtaatgtgg tcttttattg tcttaatctt tcacaaacca ttgctttcct actccaatag 300
gaagcctgat cagatactag atagctctta tatattgcca atgtatggac tgatgaaact 360
gaggctcaga gaggaagtaa aataaagcat ggctcccccc tactggttta ctatattcca 420
aagttattaa acacctggcc agcactgctg tgtaggttgt acactgcaca attgtgtccc 480
ttctaaggag tcagtgttca aatcacagac atcagagatt tattatgata atttttctgg 540
cagatggcag taaagtagct tattctaaca aaattgggaa tataaagact attttctaac 600
agatggaagt agtttttttga agcaggtgag ccttttcctg atgctcactc aggtgctaga 660
tggactagca gaaagaatgg cgtcattgat gtccctttgt atgtgttacc cagtttaatc 720
ctggggaatt ttactttttg ctggaaaagg agtcaccctc ccttgccaac cacatgtgtg 780
gttatacatt ggtattgcag agtgatgcca tttacaagta atacatttga gttggcagat 840
ttcccaaggt ttc 853

<210> 51
<211> 383
<212> DNA
<213> Homo sapiens

<400> 51
tgaaaattga ccaacaacca gaagtgacag caacaacaaa ataccaagtc agtaaagatg 60
gagagaaata gggaagcagt gaaggtagat gtcatttctg tttttagtgg tggaatacaa 120
ggtgttcttg tgcttaaagg tcatgttctt gtgataaaac gcactgcaga gacaacatag 180
tttaattggc tgaggcaggt gactcccttt aagcatcagg gtggaacaaa ctacacgaca 240
aaatgtaatt ttaaaaacca ctctcattca aatgtaagaa tatcaaagca cccttaactc 300
attacatgag tgaaacaatg agtgtcatgg tctgaattgt gttcccctcc ccaaacccgt 360
atgttcaagc cttaaaccct agt 383

<210> 52
<211> 3342
<212> DNA
<213> Homo sapiens

<400> 52
tgaaaattga ccaacaacca gaagtgacag caacaacaaa ataccaagtc agtaaagatg 60
gagagaaata gggagcagtg aaggtagatg tcatttctgt ttttagtggt ggaatacaag 120
gtgttcttgt gcttaaaggt catgttcttg tgataaaacg cactgcagag acaacatagt 180
ttaattggct gaggcagtac ctactgcacg acccccccacg tccgcctcct gccattgcca 240
gcaggtgcct tgcgccggta cctggctgcg cttattcatc cattatggtc gctctgtcac 300
tggtgccatt atgtgctcac atgcccactc cctcaggttt agaagtcgcg ttgcccggca 360
acagaacaat ctgctggctt agcctttggc caagttggca gctggacgag gacgctcaga 420
gcccagctct tgagagttca agtatccgac agttccccac tgctcccagg agcggttacc 480
cgggcactct gtgcccctca ttcctgtttg ggccaaggcc gaggacctgc gagtaccaca 540
gtaagccagt gctgtgtgct ccgagttcca gggcatcccc cagctcagcc actacactga 600
gcacaaggac tctgtggggc ccaggagcag gtagtcaccc ctttggggtc cacaacaccc 660
ggctgtcccc agacttgtgt ccagggaaga tagtgttgag ggccctcaag gagagcgggg 720
cagggatgcc tgagcagcac aaggacccca gagtccaaga aaatcctgat gatcagagaa 780
cggtccccga ggtcaccggg gatgcacggt ctgcattttg gcccctgcgg gacaatggag 840
gcccctctcc ctttgtgccc aggcccgggc ctctgcagac agacctccac gcccagagct 900
cagaaatcag atataaccac acatcccaga catcctggac gagctcgagc accaaacgaa 960
atgccatctc cagctcctac agctccacgg gaggcttgcc ggggctaaag cagaggaggg 1020
ggccagcctc atcccgctgc cagctgaccc tcagttactc aaagacagtg agtgaggaca 1080
ggcctcaggc tgtctcttcg ggtcacacac ggtgtgaaaa gggggcagat acatcaccag 1140
ggcagacaat cgccccaacg ggtggctccc ccagatccca tgactctagg ccccgtagac 1200
gcaagattcc cctgctgcca cgcaggcgag gggagccttt gatgctgcca cctcccttag 1260
agctggggta ccgggtcacg gctgaagacc tgcacctgga aaaagagacg gcattccagc 1320
gcatcaacag tgcactgcac gttgaggaca aggccatccc ggactgcaga ccctcacggc 1380
cttcccacac tttgtcctca cttgcaacag gggcttcggg tgggcctccc gtttctaaag 1440
cacccactat ggatgcacag caggacagac ccaagtccca agactgcctg ggcctagtgg 1500
ccccccctagc atctgctgca gaggtccccg ctacagctcc cgtgtctggg aagaagcaca 1560
gaccaccagg accccctgttc tcctcctcag atccccttcc tgccaactct tcccactccc 1620
gggactcagc ccaggtcacc tcgatgattc ctgcccccctt cacagctgca agcagggatg 1680
ccggcatgag aagaacaagg tcggctcctg cagctgccgc agcagcccct ccccccctcca 1740
cattgaaccc cacgtcgggg tcgctactca atgcagtgga tggaggcccc tcacatttct 1800
tggcctcagc cacagctgca gcacgtgccc agaggtcaga agtgagatat aaccagagat 1860
cccagacctc ccggaccaga tcctgcctca aacgaaatgc cagctccagc tcccacagct 1920
ctacggaagg cctccaggaa gtaaagcgga ggagggggcc agcctcatcc cactgccagc 1980
tggcccacag ttcctcaaac acagtgagtg aggacggacc tcaggctgtc tcttcgggtc 2040
accgctgtga aaacaaggca ggtacagcac cagggcagac acttgccccc aggggtggct 2100
cccccagatc ccaggcctct aggccccaca tcaacactgc actgcacgtt gaggacaagg 2160
ccatctcgga ctgcagaccc tcacggcctt cccacacttt gtcctcactt gcaacagggg 2220

cttcgggtgg gcctcccgtt tctaaagcac ccactatgga tgcacagcag gacagaccca 2280
agtcccaaga ctccctgggc ctactggccc ccctagcatc tgctgcagag gtcccctcta 2340
cagctccccgt gtctgggaag aagcacagac caccaggacc cctgttctcc tcctcagatc 2400
cccttcctgc cacctcttac cactcccggg acacagcaca ggtcacctcg ctgattcctg 2460
ccaccttcac agctgcaagc agggatgccg gcatgagaag aacaaggtcg gctcctgcag 2520
ctgccacagc agcccctccc ccctccacat tgaacaacac gtcggggtca ctactcaatg 2580
cagtggatgg aggcccctca catttcttgg cctcagccac agctgcagca cgtgcccaga 2640
ggtcagaagt gagatataac cagagatccc agacctcccg gaccagatcc tgcctcaaac 2700
gaaatgccag ctccagctcc agctcccaca gctctacgga aggcctccag gaagtaaagc 2760
ggaggagggg gccagcctca tcccactgcc agctggccca cagttcctca aacacagtga 2820
gtgaggacgg acctcaggct gtctcttcgg gtcaccgctg tgaaaacaag gcaggtacag 2880
caccagggca gacactcgcc cccaggggtg gctcccccag atcccaggcc tctaggcccc 2940
acatcaacag tgcactgcac gttgaggaca aggccatctc ggactgcaga ccctcacggc 3000
cttcccacac tttgtcctca cttgcaacag gggcttcggg tgggcctccc gtttctaaag 3060
cacccactat ggatgcacag caggacagac ccaagtccca agactgcctg ggcctactgg 3120
ccccccctagc atctgctgca gaggtcttct ctacagctcc cgtgtctggg aagaagcaca 3180
gaccaccagg acccctgttc tcctcctcag atccccttcc tgccacctct tcccactccg 3240
gggactcagc ccaggacacc tcgctgattc ctgcccccctt cacacctgca agcagggatg 3300
ccggcatcag aagaatgttt cgtgttcgaa attgtttgag gg 3342

<210> 53
<211> 129
<212> DNA
<213> Homo sapiens

<400> 53
agctgtcata cttatcgttg ctgcttatta gtatttttat ggtttgttat ttcaaaagaa 60
attcatttcc cataacaata ttattttctt aaatatgtta agctttaaaa taaaagcata 120
tcaaatgga 129

<210> 54
<211> 201
<212> DNA
<213> Homo sapiens

<400> 54
catgccgtgg cccatgccca tttgcacagg gacgcagggg gttctcacac acaggcaggg 60
tccgccccca gctgccgtcg gcgtcagtcc acacacatag gcttttgggc ggtgctggaa 120
gcttctggcc cctgaacgtt ccccccaggc cccgtttcca gggaaaggga taggcaggcg 180
cacgctgcgg ccgtttccac a 201

<210> 55
<211> 227
<212> DNA
<213> Homo sapiens

<400> 55
catgccgtgg cccatgccca tttgcacagg gacgcagggg gtctcacaca caggcagggt 60
ccgcccccag ctgccgtcgg cgtcagtcca cacacatagg cttttgggcg gtgctggaag 120
cttctggccc ctgaacgttc cccccaggcc ccgtttccag ggaaagggat aggcaggcgc 180
acgctgcggc cgtttccaca atccgacctc gtagctgggg cgtgccg 227

<210> 56
<211> 271
<212> DNA

<213> Homo sapiens

<400> 56
catcttttta atattcagta tgaccgaata aagcactggt gctgccttag taacaatggt 60
tgtnctcaag gtaaacttct catgtgcttg tttcagttgt gagctcaatt agcctctttc 120
tcatgaaatg aatgcctttt tacttgaaag aatgactgag agccaggcta tggatattca 180
aacatgtatt tttcagacac ttcttgaaaa taagtgaagc aaacctgtta attacaaggg 240
aagcaatgac aatatttgtt gccaatgata a 271

<210> 57
<211> 573
<212> DNA
<213> Homo sapiens

<400> 57
gggcaagaca gctgtgggag ttgttggttc cagtatcttt ttcttccttt ccggatggga 60
agagtgcatc tctcttggcc cacaaatgtg tttgcacact aaggtgatgg catccttaga 120
aggaagcaga gcagtgctct gaccttcgct tctggaaccg agaaaatgat gccatgctgc 180
tttgttggtg tgattgttgt tggttttttg tggatgaatt ttaaaatagt atttgtgact 240
atcatttcat gtgtccactc tttttaaaaa tgttaccttt tctaggattg gcagaatttg 300
gaattatatg tcttattaaa tatgctttga aagacagaag taataagttc tggttaatct 360
tttatagtgt ttgtcttggg agcaaatagt atgagagaga ggtgtgagaa tgggaataat 420
aatctaacat atcaaaatta gagaacccca aaccatcaca ttctttctct ttgtgccatt 480
ttagaattga gaataccgtc cttcttactg tggttatatt tttacttttg tatataaact 540
tgtagcagaa aataagattc agtagcttaa agg 573

<210> 58
<211> 843
<212> DNA
<213> Homo sapiens

<400> 58
gggcagacag ctgtgggagt tgttggttcc agtatctttt tcttcctttc cggatgggaa 60
gagtgcatct ctcttggccc acaaatgtgt ttgcacacta aggtgatggc atccttagaa 120
ggaagcagag cagtgctctg accttcgctt ctggaaccga gaaaatgatg ccatgctgct 180
ttgttggtgt gattgttgtt ggttttttgt ggatgaattt taaaatagta tttgtgacta 240
tcatttcatg tgtccactct ttttaaaaat gttacctttt ctaggattgg cagaatttgg 300
aattatatgt cttattaaat atgctttgaa agacagaagt aataagttct ggttaatctt 360
ttatagtgtt tgtcttggga gcaaatagta tgagagagag gtgtgagaat gggaataata 420
atctaacata tcaaaattag agaaccccaa accatcacat tctttctctt tgtgccattt 480
tagaattgag aataccgtcc ttcttactgt ggttatattt ttacttttgt atataaactt 540
gtagcagaaa ataagattca gtagcttaaa ggggccaggc actgtggctc atgcccgtaa 600
tcccagtggt ttagaaggct gaggggaagg atcacttgag gcctggagtt tgaaatcaac 660
ctgggcaaca tagcaagact ctgttccttc aaaaaaaaat ttttaaaaat tagctgagca 720
tggaggtgca tgcctataat cctagcaatg attataccat tacactccag cctggatgat 780
agagtgacac cctgtctcaa agagaaaaaa aaaaaaaaaa aattctgcgg cgcaagaatt 840
cgc 843

<210> 59
<211> 221
<212> DNA
<213> Homo sapiens

<400> 59
cgggactgga agggtgtgcc ctcgccgtcc tcgccttcgt cttgcacggg acaagatgtc 60
acgattccga atccaaacct cagagacagc cccatccct ctcgttagcc acccacacac 120

cccgctcagc aacaataaca acctgcattt agggaacgtg tgttatgtgc caggccacac 180
aggcattatc tcatgtactc ctcacaggca ccttatcaag g 221

<210> 60
<211> 535
<212> DNA
<213> Homo sapiens

<400> 60
gcacgtgggg tcggggtggg ggcgaagggc cgcttggcct ctgtagggtc gggactggaa 60
gggtgtgccc tcgccgtcct cgccttcgtc ttgcacggga caagatgtca cgattccgaa 120
tccaaacctc agagacagcc cccatccctc tcgttagcca cccacacacc ccgctcagca 180
acaataacaa cctgcattta gggaacgtgt gttatgtgcc aggccacaca ggcattatct 240
catgtactcc tcacaggcac cttatcaagg agatgctgtt gttacctgca ttttacagat 300
ggggaaactg aggctcaatg cattaaggac tgccaggaag ccctgtcctg tggctgtgat 360
gatggaaatg ttccctgtgt tgttcagtat ggtagtcact ggccacaagt gagcactgga 420
aatgtgccta ttgagactga ggaactgatt ttttcatttt gtttaattgt aattaaacag 480
ttacgtgtgg ctgtggtatt ggaaaaaaaa aaaacaaaaa aaaaaaaaaa aactc 535

<210> 61
<211> 514
<212> DNA
<213> Homo sapiens

<400> 61
cgtctcactg atatcccctt cagttctccc aaatcaccct ttctgaaaca tacatccgat 60
catgtcattc ctttgcaaaa ctaagtttcc ctttgcactc aaaacaatat ctgaatgtct 120
tgctctggtt tctcaggccc cgcctctacc actggcctca gctcttcccc tctctccatt 180
gctcactgaa taacagccac caagacctcc ttgccattgc tcaaacatgc aaggcctaca 240
cctgccacag ggccttggca catgctattc catctgttta caatgcttgt ctccacatgg 300
ctacttcttt gtagcagttg gtctcagctc aaatgtcatg tccccaacca gcctacctaa 360
agcagttctc cctacctagg cctttcttgc tcaccatgta aaagattcct atttagtttc 420
tgttattatc cttcttgctc tagaatggaa gccctacgag ggcaagatat ttttctgtat 480
cggtcactgc tatagcttca acaccaagaa catg 514

<210> 62
<211> 598
<212> DNA
<213> Homo sapiens

<400> 62
tttcttctgg tcttattcct ctggatctaa atatatcaac attgacatgt gaggtcagaa 60
ttttatcatt tgaaaaattt ttttcacttt actgaaaatt cagtcacttc ttcctgaaga 120
aggaagcaat tgcatatcaa ttttcttcta ttccagctta atctatttat ttttctcttt 180
tacattaaaa cattctttta atgatatatg ctgcctgtaa atatttccca cccactttcc 240
agaggtaatc cactgttatc aagtaagttt agtaaatttt ttttaattga atttctcaa 300
taggtcatta acgtgtttca aagttgaaaa attacaaaac tatgtgtcgt gaaaagtctc 360
cttctttccc ttgtgtccca agctacctag ttcttggagc cagttgatgt tatcagattc 420
tttggtattc tttcagacac acatggtatg cattatttga gcaaaggggc gtgggtgtgt 480
gtccctctgt ttttaagttc taaatgttag catgctacac atactttttt catatatttt 540
cttaagtaac ttcatttcat tatatgtatt ccactttgta aaattagata ctacatgc 598

<210> 63
<211> 648
<212> DNA

<213> Homo sapiens

<400> 63
tgaatttctt ctggtcttat tcctctggat ctaaatatat caacattgac atgtgaggtc 60
agaattttat catttgaaaa attttttca ctttactgaa aattcagtca cttcttcctg 120
aagaaggaag caattgcata tcaattttct tctattccag cttaatctat ttatttttct 180
cttttacatt aaaacattct tttaatgata tatgctgcct gtaaatattt cccacccact 240
ttccagaggt aatccactgt tatcaagtaa gtttagtaaa ttttttttaa ttgaattttc 300
tcaataggtc attaacgtgt ttcaaagttg aaaaattaca aaactatgtg tcgtgaaaag 360
tctccttctt tcccttgtgt cccaagctac ctagttcttg gagccagttg atgttatcag 420
attcttttgt attctttcag acacacatgg tatgcatttt tgagcaaagg ggcgtgggtg 480
tgtgtccctc tgtttttaag ttctaaatgt tagcatgcta cacatacttt tttcatatat 540
tttcttaagt aactttattt cattatttgt attcagtttt gtaaaattag atactacatg 600
catgtggttc aaaagtaaaa tgatgtaaag gctaataata tgtaatag 648

<210> 64
<211> 601
<212> DNA
<213> Homo sapiens

<400> 64
gaggtaagcc cattctcacc ctacaggcag ggtagggagg ggcacattct gagaagtgga 60
tatcacaagt tgtccctcat gtggatccca tgacttgagt agtgggaaac tccaagctga 120
tttgaagatt ctttatttga gagtggacct gcacacctag tgttcctggg tcagtccagg 180
ggcgagcaga tcattgaagg actgcacctt tatcctaggc tcaccaaatc cccaggtgta 240
attatcaaac gaaagagaat agtccatagt taggaacaac taaatgcaaa ggatgaaata 300
gcactgaatg aggaccagca ggaagagatc tcagaaaaca taagataatg gacttgttga 360
catagatttt aaaggcctta ctcaaactaa ataggcaaga ttgaaaatac ctatgaggaa 420
gaggaaacta tacaatgacc tagcaaattt gaaaaaggaa ccaggaacaa cttgtagaca 480
tgaaaagttc atgtctttat aataaaaatc taacagatgg atttactagc agattacata 540
aaactgaaga gagtgaatga cctggaaagt agagaagaag aaatataatt tagagaacca 600
c 601

<210> 65
<211> 1216
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (58)..(125)
<223> a, c, g or t

<220>
<221> unsure
<222> (1204)
<223> a, c, g or t

<220>
<221> unsure
<222> (1206)
<223> a, c, g or t

<400> 65
tctcctgatc ttggacatca gaactcctga ttctcaagcc tttgggtttg gactggannn 60
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 120
nnnnntaact ctaatacacc agtacaatgg aaagattcct aaattcaaaa gccagaaggc 180

```
tgggttcctg ttcccaccct gccttttacc ttctgtgtgt tcctgatgaa gacacttcat 240
gctccactat ttacttacct ctgaaacgaa gggctgaccc agatcagttg ttctctgacc 300
tgcttggagg gactcagagg ctgtggagag tgaatatttc catcagctga tgcccttctc 360
aggcctcaat ctcccctggt ttgcagactg tggccctcct tggcctctgt ggaatctggc 420
catttgaatc ctgtcagccc tgttttccat caccaaagga ctccggagga actgtgccaa 480
gcaggctgtc agccacgggc aagctttctg aaaaagacgt gccaactgca gccacagaaa 540
gtccattccc ttgaataact ctgctaatat ttgaaaatta gttcccttgc tcctgatcat 600
gctactgggt atttggatat aagagccaag gatgagggca atagaaaatt aaaatcatgt 660
tctactcata taaactgcac agatatggaa gggtaggtcc tattacctat aatcctggga 720
tttttagact ctcactttca ttggaccaga gttgccttag ggacagtaaa aacacaaaat 780
gctgggtatt gttttcatca agcaactact gatagtgcac atttaaatca aaattcttct 840
aatcccaaac tcagtaaaca gatgctgtga gcttagttct gcccctctgg cttcagattt 900
taccccactg gatgtgccca attctgagat gacaagacgc ttccagcttc cacatggttg 960
caatttggct gtggaactgg catgaaagca cgtcactgtg tcagcacctg ggccaccaga 1020
tgaataacct atgaacaaca gctttggact aaaatatgaa ggggttgttt tccttcaatc 1080
tccccctacc ttcctcagaa cctgctacaa ggaaagattt atagactcga aagcgtcaat 1140
gactgattag acccatatga ttgctcctgc tgtttctgat attttaaaaa attgtcttat 1200
aaangnataa aaataa                                                 1216
```

<210> 66
<211> 1430
<212> DNA
<213> Homo sapiens

<400> 66

```
gctcaccaat gtgggtgggc ctcattcaat ccattgcaga cttgaataga actaaaagga 60
agaggaaggg caaatttgtt ggctgcttga gctgggatat tcatctttct cctgatcttg 120
gacatcagaa ctcctgattc tcaagccttt gggtttggac tggaggcacc agctttcctg 180
ggcctccagc ttgcagatgg catatcatgg aacttctcag cctccaaatt cataactcta 240
atacaccagt acaatggaaa gattcctaaa ttcaaaagcc agaaggctgg gttcctgttc 300
ccaccctgcc ttttaccttc tgtgtgttcc tgatgaagac acttcatgct ccactattta 360
cttacctctg aaacgaaggg ctgacccaga tcagttgttc tctgacctgc ttggagggac 420
tcagaggctg tggagagtga atatttccat cagctgatgc ccttctcagg cctcaatctc 480
ccctggtttg cagactgtgg ccctccttgg cctctgtgga atctggccat ttgaatcctg 540
tcagccctgt tttccatcac caaaggactc cggaggaact gtgccaagca ggctgtcagc 600
cacgggcaag ctttctgaaa aagacgtgcc aactgcagcc acagaaagtc cattcccttg 660
aataactctg ctaatatttg aaaattagtt cccttgctcc tgatcatgct actgggtatt 720
tggatataag agccaaggat gagggcaata gaaaattaaa atcatgttct actcatataa 780
actgcacaga tatggaaggg taggtcctat tacctataat cctgggattt ttagactctc 840
actttcattg gaccagagtt gccttaggga cagtaaaaac acaaaatgct gggtattgtt 900
ttcatcaagc aactactgat agtgcacatt taaatcaaaa ttcttctaat cccaaactca 960
gtaaacagat gctgtgagct tagttctgcc cctctggctt cagattttac cccactggat 1020
gtgcccaatt ctgagatgac aagacgcttc cagcttccac atggttgcaa tttggctgtg 1080
gaactggcat gaaagcacgt cactgtgtca gcacctgggc caccagatga ataacctatg 1140
aacaacagct ttggactaaa atatgaaggg gttgttttcc ttcaatctcc ccctaccttc 1200
ctcagaacct gctacaagga aagatttata gactcgaaag cgtcaatgac tgattagacc 1260
catatgattg ctcctgctgt ttctgatatt ttaaaaaatt gtctcataaa gagatacaaa 1320
taaataatca atggcaaact tctggcatgg gagagacatt tagggaaaga agtcatctca 1380
gcctccccca acacacacat gcacacacac atacagctgc aaacacaatt           1430
```

<210> 67
<211> 430
<212> DNA
<213> Homo sapiens

<220>
<221> unsure

<222> (72)..(139)
<223> a, c, g or t

<400> 67
gggatattca tctttctcct gatcttggac atcagaactc ctgattctca agcctttggg 60
tttggcctgg annnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 120
nnnnnnnnnn nnnnnnnnnt aactctaata caccagtaca atggacagat tcctaaattc 180
taaagccaga aggctgggtt cctgttccca ccctgccttt taccttctgt gtgttcctga 240
tgaagacact tcatgctcca ctatgtactt acctctgaaa cgaagggctg acccagatca 300
gttgttctct gacctgcttg gagggactca gaggctgtgg agactgtggc cctccttggc 360
ctctgtggaa tctggccttt gaatcctgtc agccctgttc tccatcacca aaggaatccg 420
gaggaactgt 430

<210> 68
<211> 829
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (240)..(354)
<223> a, c, g or t

<400> 68
gtatgtacta cccacataag tgggacactt tgaacaatga aatatagatg ttttcaccaa 60
agaagggagt cttatttttt tccgacttca gacaattcat cttcatccat taatttttcc 120
tttttgtaat atgtaccttt atgctaattt ttaatatgca aataacttac aaatatatgc 180
tcagcatttg agtacaggct gtgctttatt acatattaca tgcatgtatg caatgtactn 240
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 300
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnacaaaa 360
taaaatttgg aatgaagcag gaattatctt ggactattta taatttatta agatactaaa 420
taccgtcatt atgaaatggt ctcattaagt gatccctgtc taaagagttg cataatagtg 480
agacaataag gggcttagtg tatttttttt cttttgaaca taagctattg tacatttgtg 540
ccaacaggac ttctttatag agtctcattt tcctattaca atattatttt tgttattaag 600
tgaaacacct catatcacca ccactgctga gccagatata atagactgta ctgtgtaagg 660
ttcttaaaac tcacatctat aataaccaga cctctttttt tatattgatt caaattatgt 720
ttaatgctga attataagca aaacctacaa gaataaaatc attttatgct ttgaaactga 780
ctcctttttt aaaaaaagaa tgatcacaac taccaactcc ctcatctat 829

<210> 69
<211> 541
<212> DNA
<213> Homo sapiens

<400> 69
atagactagt aaagtctgtt tttatataaa agtgacacag gaagctgtta caatctagga 60
atgggcaggt atggtcagtg gttgtcacaa tagagccacc caaggagaca tctcttctcc 120
agatcctaac agagtgcatc ttgtgctttt cctaacagac ctgtcggact ggctttttct 180
cttttaagga tatagagaaa gcaaaattag caaatctagt ttcttgtcac tttactagga 240
gggaggaaaa gagagaaaga atgcacttgg gaatgggagg ccttgctttt aatttaccag 300
atgccagtta gagcgttaat gccacacgag ccagagaggt caccttgctg agcatggctt 360
gactgttgca gcctctttct gcgactccag acatgcgatg tctgttagct gattctagcc 420
ttcagatgca gcccggagat gtaaccctga ggctggagtc ctgtggctct aatcccagac 480
agaggcaact ccaccaagtt ctggtttggg tcagaaatag agggaaagga tgaatgaaag 540
a 541

<210> 70
<211> 696
<212> DNA
<213> Homo sapiens

<400> 70
atagactagt aaagtctgtt tttatataaa agtgacacag gaagctgtta caatctagga 60
atgggcaggt atggtcagtg gttgtcacaa tagagccacc caaggagaca tctcttctcc 120
agatcctaac agagtgcatc ttgtgctttt cctaacagac ctgtcggact ggctttttct 180
cttttaagga tatagagaaa gcaaaattag caaatctagt ttcttgtcac tttactagga 240
gggaggaaaa gagagaaaga atgcacttgg gaatgggagg ccttgctttt aatttaccag 300
atgccagtta gagcgttaat gccacacgag ccagagaggt caccttgctg agcatggctt 360
gactgttgca gcctctttct gcgactccag acatgcgatg tctgttagct gattctagcc 420
ttcagatgca gcccggagat gtaaccctga ggctggagtc ctgtggctct aatcccagac 480
agaggcaact ccaccaagtt ctggtttggg tcagaaatag agggaaagga tgaatgaaag 540
aagatacaaa gaaataatga acaagtgagt tctttcagct gcttacttgg gtggtctgca 600
ggcagcaaga gacaggaagg aggctgttgt ggggtccttg ttcgaggcag tgggagattt 660
gctcagaggg gttgtgtggg aagtgagaga aggggt 696

<210> 71
<211> 1207
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (957)
<223> a, c, g or t

<400> 71
gcagtgccag gacctctccc ggaggcgggg cagagcagca gcttctcggc cctgtgccga 60
gcccaggcct gcacccctaa ggcaggcact gctccgtgat ccaggaacca cctctctcta 120
cagctgggag tgagcagtca gagagggaga cagccttgcc cggtgctacc cagcaagcta 180
gtcaccgagt gggcagaggg aggagcggcc ctcaccggat gtcaagcagc ctgggtcccc 240
agtccagctc tgcctgtccc tcgcaataac gcctcagtga cgaccatttg tgagccatct 300
ctctgtctca ggcacggtgc tacatgccaa cgaaacctgc tcccattgaa ccctggccag 360
ccagtgaaga aagggttggg cctgggaggt gccactttac agacaggggc accaaggggc 420
agggtggcag gaggcccacc ggacgttccc catgaagtag cagtcccagc atccacaccc 480
agcaggcacc acgctggccc gcagcctccc tgccagcacg cctggcttcc cggcctcgga 540
acttgatctg ctccctcttc cggacactgg ggctcctgcc aagtcctggg ctgggcagca 600
actgctgaac attctaagaa atccctccca gggttttctc aggagcccgg gtggggcagg 660
aagtccccag gggctgaggg gaccgtggcg gcaggtggca cccagagcag cactctcctg 720
gggcccaggc tgttgggcca gaggcaggac tgtgaggcct agtgtagggc ctcctgccag 780
tggccggcac ctacttgtgg ggctgggggt tcccccagca ggttgggctc cccacctgac 840
acactcacag accttgtgcc ttggagagcc agtgttcccg gggccacata gctatgccgc 900
ccaggggctg ggcctgtccc agctctggtc ccccggcccc aggtcctgga cgctggntcc 960
gcgcagcagc aggcggcctc cggaggacac gatgtgactg gctgccgcta cgtcgcactc 1020
agatgagtct gcgccggatc gacctgctgc cgagtcctgc cggacaggca caggcaggga 1080
gtgaaaatta tctacccctt tttatttctt aataactgaa tgaaaataaa cattggtggt 1140
ttgacaaata actacatatt ttcaaaccca gccagtccag gggatgcagt ttccaggtgc 1200
gttatgc 1207

<210> 72
<211> 263
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (231)
<223> a, c, g or t

<220>
<221> unsure
<222> (239)
<223> a, c, g or t

<220>
<221> unsure
<222> (242)
<223> a, c, g or t

<220>
<221> unsure
<222> (248)
<223> a, c, g or t

<220>
<221> unsure
<222> (259)
<223> a, c, g or t

<400> 72
gtcctacttc aataatttaa aaaaatattc tgggatttgc attcctcaaa tttcagccct 60
cattttactt tacctgtcta cagtgttttg cgcaattgac cactccttcc tttttgaagt 120
attttctttc cttggtttct gaaatactgt tatcttccta tctcactggc catacattct 180
agtctccttt gctagtttat tatggttttc atcttctcaa caacaatttt nttttttttng 240
gnggagangg agtcttgcna tgt 263

<210> 73
<211> 579
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (547)
<223> a, c, g or t

<220>
<221> unsure
<222> (555)
<223> a, c, g or t

<220>
<221> unsure
<222> (558)
<223> a, c, g or t

<220>
<221> unsure
<222> (564)
<223> a, c, g or t

<220>

<221> unsure
<222> (575)
<223> a, c, g or t

<400> 73
tggtgtttga gtcctaaaaa ttgtatacca gtgctaattg ggatatacca aacttttttgt 60
ctcttagtaa ttagttttgt tttgttttgt tttgttttaa tgttgtgctt atcttaaggt 120
gtaaatgcag acaaagttgg aattgaagct gccgaaatgc tattagcaaa tcttagacat 180
ggtggtactg tggatgagta tctgcaagac caggtaatga cacatttagg ttaaaaaccc 240
tctaacctgt tagatttgaa tatgtggtag attgaatatc aatttaaata attgactttc 300
agacactaat tagcaagtcc tacttcaata atttaaaaaa atattctggg atttgcattc 360
ctcaaatttc agccctcatt ttactttacc tgtctacagt gttttgcgca attgaccact 420
ccttcctttt tgaagtattt tctttccttg gtttctgaaa tactgttatc ttcctatctc 480
actggccata cattctagtc tcctttgcta gtttattatg gttttcatct tctcaacaac 540
aattttnttt ttttnggngg aganggagtc ttgcnatgt 579

<210> 74
<211> 339
<212> DNA
<213> Homo sapiens

<400> 74
ctctgttcct tgctcatctt catggtgatt gggggtagat cagatgagtg tgtaaaagcc 60
ccttgaaagc tggaaagagc ttaacaaata tcagctgttg ccatgaaaga atatttgctt 120
actttccatt gtgtataaga taacgataat catagaatta atattattca acttccttgt 180
gtcttttgca catttctgta cagtcctgtt tttgtttgtt actgtcattc tcaaagtact 240
caagttgaat tttgtcactt tggatttctt ccaggaatat gtgagagaca tttaggtctc 300
taatgatgaa gtattttcta ggcgtaatgc aaaagattg 339

<210> 75
<211> 299
<212> DNA
<213> Homo sapiens

<400> 75
caacgacaga taacttcgtg atggaaaatg taggtctcct tagtagttag ccctctgcca 60
ggtgacttcg tttccacctc ccctatata ttgttcttcc ttcctctcta aattctctaa 120
atctctgctt atacagagca atctggctct ctctggcctc tccagtcatc atacatcata 180
ctcacattca ccatcttgag aagtgcagta agccacataa atgcagcaga agtaccttat 240
gcagtcctag gaggctgtgg ttttgagttg cttttttttt tcttttggga gacggagcc 299

<210> 76
<211> 247
<212> DNA
<213> Homo sapiens

<400> 76
tgtatattga gctcctactg tgtggcaagg cctatggtaa gcattttatt ttggtaactt 60
gtttaatcct cattacaatt ctgtggtaaa tgctattatc tgtttttata ttgaagggat 120
gaaatggagg ctcagaggga tatgtagtag ctaaatgtta gagctaggat tganacccaa 180
attgacttct gagtatagat ttccccccaa ctgtatgata cttcatattt ggagtcagct 240
tgaagta 247

<210> 77
<211> 254

<212> DNA
<213> Homo sapiens

<400> 77
tgtatatttg agctcctact gtgtggcaag gcctatggta agcattttat tttggtaact 60
tgtttaatcc tcattacaat tctgtggtaa atgctattat ctgtttttat attgaaggga 120
tgaaatggag gctcagaggg atatgtagta gctaaatgtt agagctagga ttgaaaccca 180
aattgacttc tgagtataga tttcccccca actgtatgat acttcatatt tggagtcagc 240
ttgaagtaat tcac 254

<210> 78
<211> 504
<212> DNA
<213> Homo sapiens

<400> 78
tgatttgatt tggttttaaa atagaatagt tgtactctga gggaggaggg aaatgcttaa 60
acaatactaa gaattccatt ctttagagac aaattactta gaagttgata gtgacatatt 120
gaaagggttg ttgattgttg gattattcag gtgatgaaga tgatggtagg ggccatggcg 180
gctgagggag aatgagtctt aaacactgag gaggcacaaa agattgggtg gctggatata 240
ataggaaact ggaacgaaag aaggagaaga gaatggcgat actgataaaa aatagaatga 300
aagaagatgt gtggaaaaga aagtttcact ttgaaggctt gatttttgaa gtgatggcag 360
atatagatat acatccaata gatgagtggg aaaagtaaat caaacagaaa tgaaaaattg 420
agtccaagat tgatgggaga ctaataatgg ggaggactga gcctgggggc aactacatta 480
gtaacagtgg caggttttgt tttt 504

<210> 79
<211> 210
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (80)..(99)
<223> a, c, g or t

<220>
<221> unsure
<222> (173)
<223> a, c, g or t

<220>
<221> unsure
<222> (175)
<223> a, c, g or t

<220>
<221> unsure
<222> (206)
<223> a, c, g or t

<400> 79
gtccctctag agaaccctga ctaatacagg tggttcctgg ctcatggcag tgtgactcca 60
gtctttacat ggcgttcccn nnnnnnnnnn nnnnnnnnnc aaatttcctc ttttcataag 120
gaccgtggta ttggataggg gtccacccta cttcgatatg accttatttt aantncatct 180
ttgatgaccc tgtttccacg taaggncaca 210

<210> 80
<211> 161
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (116)
<223> a, c, g or t

<220>
<221> unsure
<222> (148)
<223> a, c, g or t

<400> 80
gagggtcaga agcagaaaga tgacatcata agaaagactc aactggccat ttttggcttt 60
gaaggtggaa aggggacctg agtccaggca tgtgggcagc ctggagaagg cgaganaatg 120
gattcttccc cagaatccct ggaaaggnac gtggccctaa c 161

<210> 81
<211> 112
<212> DNA
<213> Homo sapiens

<400> 81
tagcaccttt taataactct ttttagagta atttagagca aactagataa attttaatat 60
atatctcatt gcatactttt atgtaacttt gtcttagaaa aacaagagtt ct 112

<210> 82
<211> 277
<212> DNA
<213> Homo sapiens

<400> 82
tgaaatgatg acaccagtag aatatggtga gatatgtata cacaatgtaa tacctagagt 60
gacaatttaa aaacctatac aaagagtgac acataaataa acaaaaacaa cataaaaata 120
aaaatataat tctaaaaata ttcaagtagc caattggaag gtggaaaaaa gaaaaagaac 180
aaaaaataga acagcactaa acaaaaaata aaatcgcaga cctaggccct gacatatcaa 240
taattatatt aacatgtaaa tggtctaaat tttacca 277

<210> 83
<211> 637
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (92)..(196)
<223> a, c, g or t

<220>
<221> unsure
<222> (230)..(316)
<223> a, c, g or t

<220>
<221> unsure
<222> (367)..(428)
<223> a, c, g or t

<400> 83
gtccttnggt gttgcaccaa acaggctaag aagcaatgac attgattatg aggaacttgg 60
aactcagatg tattaatttc ctattgtgtc tnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 120
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 180
nnnnnnnnnn nnnnnncact ttctttctgt aggctctagg agagaatcta gnnnnnnnnn 240
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 300
nnnnnnnnnn nnnnnnccaa gtccttctca cactgctgtc tttttggttc tctctcttgc 360
ctgcctnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 420
nnnnnnncat agttgattag cagccttaat ccatctgtaa ttttaattcc cttttgccag 480
gtaatgtggc cattatcttg cctacaacct cagaggatgt tgataatgta aagggtagtg 540
aattggggag ttcatagggt ttgatagttg acaaatacag agtgtagtat taggtagggg 600
tttttggca gggtgcagtg gcccatacct gtaatgt 637

<210> 84
<211> 577
<212> DNA
<213> Homo sapiens

<400> 84
caggcataag ccaccacacc tagccaagaa accattcttt gaacacaagc aaatatactt 60
tggagaaaaa tttaataatc ctggcagggc tacattcaac ataattctgt tatgggggaa 120
ggcagcatgc tttggctgct cagtgagcta tgttctgtac aaccaagtga aattgctaaa 180
aaaagattct cctgtataca gtaacttaaa gtgatgcagt ctacttaaga tcagatctga 240
gttacaaaat caaaagtgac agctcctatg ttcttttaaa gtccaatctc tttttttcat 300
tgttgtgctc caaatgcctt gagtacctga tgtagagtag gtggctaata aatattggtt 360
gaatttcttg aacgaatctg ttatgaaaag atctactttg ctcatctctg tgccccaata 420
gcaggagctt gaggagaagg agaaaatatt gggtcagagc ttttgattaa tatgtatgat 480
tctattaaac gggttcacta aaccaaaaaa ggcaaggaaa acagttaaac caagagtctt 540
gaggttcaag tcttgtgatg attaaatcat catccta 577

<210> 85
<211> 687
<212> DNA
<213> Homo sapiens

<400> 85
ttcccatgtt agccagggct ggtcttgcac tctgaacctc agggtatcac ccccccttgg 60
gctcagggct catggctgat attacaggca taagccacca cacctagcca agaaaccatt 120
ctttgaacac aagcaaatat actttggaga aaaatttaat aatcctggca gggctacatt 180
caacataatt ctgttatggg ggaaggcagc atgctttggc tgctcagtga gctatgttct 240
gtacaaccaa gtgaaattgc taaaaaaaga ttctcctgta tacagtaact taaagtgatg 300
cagtctactt aagatcagat ctgagttaca aaatcaaaag tgacagctcc tatgttcttt 360
taaagtccaa tctctttttt tcattgttgt gctccaaatg ccttgagtac ctgatgtaga 420
gtaggtggct aataaatatt ggttgaattt cttgaacgaa tctgttatga aaagatctac 480
tttgctcatc tctgtgcccc aatagcagga gcttgaggag aaggagaaaa tattgggtca 540
gagcttttga ttaatatgta tgattctatt aaacgggttc actaaaccaa aaaaggcaaa 600
ggaaaacagt taaaccaaga gttcttgagg ttaaagtctt gtgatgatta aatcatcat 660
cctaagatga tgatgacata aactttc 687

<210> 86

<211> 77
<212> DNA
<213> Homo sapiens

<400> 86
tgtcgtttta caactgtcgt gactggtgaa aaccctgtgc gttaccctaa cttaatctgc 60
cttgcatgca catcccc 77

<210> 87
<211> 575
<212> DNA
<213> Homo sapiens

<400> 87
cacacacgtg cacacacaat actcacttaa caaacattta atttattgaa catttattat 60
atgccaaagc tggtataaga caccaaaaga gtaagacaga aagtattctt ccctggagct 120
ttgtctgact ttccaagctt tattaggcat caaacaaaac tgaagtgctt tttaagattc 180
aagtctccta cgtcgtctaa ggcagagtaa gtagccttca gtactatatt ttactctaat 240
tttttttaa cacaatggca gtactataag tatgaaactt tggtataaat gtcagattct 300
agattgtgct cctgctttct gcacactcta atatttttaa acatctcgaa aatacagagt 360
ggcagcaaaa ttacctgtaa aaacatacta gctcaagagt ttgacaggct caaaataaat 420
taccttaaat acattaaaca agaagtgtat ttgttataca gtatgtactg accaaaatta 480
aagtgcaggt tgtacagaaa gagctgcttg tgttatttta tgagcaaaat gaaaagctaa 540
tttggtacat ttaaaaatca gcatctagca aattc 575

<210> 88
<211> 663
<212> DNA
<213> Homo sapiens

<400> 88
cagtaattcg gcacgaggcg cacttttttt tttttttttt tataaaaaca gtcaacactt 60
gccccaccct actcccagca tatgcacaca cacacgtgca cacacaatac tcacttaaca 120
aacatttaat ttattgaaca tttattatat gccaaagctg gtataagaca ccaaaagagt 180
aagacagaaa gtattcttcc ctggagcttt gtctgacttt ccaagcttta ttaggcatca 240
aacaaaactg aagtgctttt taagattcaa gtctcctacg tcgtctaagg cagagtaagt 300
agccttcagt actatatttt actctaattt ttttttaaca caatggcagt actataagta 360
tgaaactttg gtataaatgt cagattctag attgtgctcc tgctttctgc acactctaat 420
atttttaaac atctcgaaaa tacagagtgg cagcaaaatt acctgtaaaa acatactagc 480
tcaagagttt gacaggctca aaataaatta ccttaaatac attaaacaag aagtgtattt 540
gttatacagt atgtactgac caaaattaaa gtgcaggttg tacagaaaga gctgcttgtg 600
ttattttatg agcaaaatga aaagctaatt tggtacattt aaaaatcagc atctagcaaa 660
ttc 663

<210> 89
<211> 80
<212> DNA
<213> Homo sapiens

<400> 89
gattggatgg tgtttcagaa aacaagcctc tattcaaata atattttact ataattcttg 60
ttaaaaatac tgtatactaa 80

<210> 90
<211> 496

<212> DNA
<213> Homo sapiens

<400> 90
gccgactttt tttttttttt tttttgtatt tttagtagag acggggtttc aacatgttgg 60
ccaggatggt cgtgatctcc tgacctcgtg atccgctgcc ttggtctccc aaagtgctgg 120
aattacaagc gtgacgcact gtgcccagct tagtatacag tatttttaac aagaattata 180
gtaaaatatt atttgaatag aggcttgttt tctgaaacac catccaatct gaaagtagaa 240
gaaaaaggct gggtgtggtg gctcatgcct gtaaccccag cactttggga agctgaggcg 300
ggcggatccc ttgagctcag tttgagacca ggctgggcaa ctccatcttt accaaaaaat 360
acaaaaatga gccaggcatg gtggtgtaca cctgtggtcc cagcggctct gggggctgag 420
gtgggaggaa ggcttgggcc taggaggtgg aggttgcagt gagccaggat tgtgccactg 480
ccgatagagc cagata 496

<210> 91
<211> 385
<212> DNA
<213> Homo sapiens

<400> 91
gaaatgggtc cggacaggtt aaaacaaaaa tccaatactg ccgtagtttc taggtggata 60
taacattttt agaaatctta taatacaata ttaacttcat tggctgaacc caagcctttc 120
agcctttata gatttgccat gatcctaata catataagca ttcattgtat tcattattaa 180
ttacttcata gattcagtgt gtgacgaagg gagatgattt ttaacaaata ataaagtgaa 240
atgatctagt tttgctatgt tgnttgagca acatcaaata gttttgctaa aatagataat 300
ttatagtgat tttttttca ctatggnatt ttcttaaata tattaagggc tttcattttc 360
tgataccacc tagtttaatt ggggg 385

<210> 92
<211> 500
<212> DNA
<213> Homo sapiens

<400> 92
gaaatgggtc cggacaggtt aaaacaaaaa tccaatactg ccgtagtttc taggtggata 60
taacattttt agaaatctta taatacaata ttaacttcat tggctgaacc caagcctttc 120
agcctttata gatttgccat gatcctaata catataagca ttcattgtat tcattattaa 180
ttacttcata gattcagtgt gtgacgaagg gagatgattt ttaacaaata ataaagtgaa 240
atgatctagt tttgctatgt tgtttgagca acatcaaata gttttgctta aaatagataa 300
tttatagtga tttttttttc actatggtat tttcttaaat atattaagtg cttttcattt 360
tctgatacca cctagtttaa ttgggggtga atatcagaga aattagaatg ttatttcagc 420
tgaaggagta cagttttttt tttctcttct tagagaatat agtgcctcag atacagtcca 480
caacaaaaat tttggtttag 500

<210> 93
<211> 364
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (19)
<223> a, c, g or t

<220>
<221> unsure

<222> (21)
<223> a, c, g or t

<220>
<221> unsure
<222> (35)
<223> a, c, g or t

<220>
<221> unsure
<222> (40)
<223> a, c, g or t

<220>
<221> unsure
<222> (60)
<223> a, c, g or t

<220>
<221> unsure
<222> (70)
<223> a, c, g or t

<220>
<221> unsure
<222> (92)..(93)
<223> a, c, g or t

<220>
<221> unsure
<222> (95)
<223> a, c, g or t

<220>
<221> unsure
<222> (97)
<223> a, c, g or t

<220>
<221> unsure
<222> (121)..(122)
<223> a, c, g or t

<220>
<221> unsure
<222> (131)
<223> a, c, g or t

<220>
<221> unsure
<222> (148)
<223> a, c, g or t

<400> 93
ttaggtccaa actcaggtna ncaaaactag tcccntcagn ccaattgctt gtacatttcn 60
acaggcaccn ctttggcaaa cccacatgga tnntnanaaa tggagaatga ggagacagcc 120
nnatattaat nagatgtatc aaactgtnac aatatgtgaa gagtattgtg tatatacaaa 180
caggaaacaa ttgaaagcct tcaacatgtg tgggtggggg gagagataac tgaattaaca 240
ggccatgtag taaaacttaa aatcaaatcc agtagtcttg aaggtatagt aattgtttag 300

ttttgaaggt atagtaatta agtactgcgc actaaaaaaa actgaccaaa aggccgggtg 360
cggt 364

<210> 94
<211> 1646
<212> DNA
<213> Homo sapiens

<400> 94
ttaggtccaa actcaggtna ncaaaactag tcccntcagn ccaattgctt gtacatttcn 60
acaggcaccn ctttggcaaa cccacatgga tnntnanaaa tggagaatga ggagacagcc 120
nnatattaat nagatgtatc aaactgtnac aatatgtgaa gagtattgtg tatatacaaa 180
caggaaacaa ttgaaagcct tcaacatgtg tgggtggggg gagagataac tgaattaaca 240
ggccatgtag taaaacttaa aatcaaatcc agtagtcttg aaggtatagt aattgtttag 300
ttttgaaggt atagtaatta agtactgcgc actaaaaaaa actgaccaaa aggccgggtg 360
cggtggctca cgcctgtaat cccagcactt tgggaggccg aggcgggcgg atcacctgag 420
gtcaggagtt cgagaccagc ctggccaaca tggtgaaacc ccgtctctac taaaaacaca 480
aaaattagcc gggcatggtg gctcacgcct gtaatcccag cactctggga ggccaagaac 540
atgctgatgg tttatgccac aagttgacaa ctgtgtgtcc aactgtgaaa cctcagactc 600
aaggtctagc aaaagatgct tgggaaatcc ctcgagaatc tttgcgacta gaggttaaac 660
taggacaagg atgtttcggc gaagtgtgga tgggaacatg gaatggaacc acgaaagtag 720
caatcaaaac actaaaacca ggtacaatga tgccagaagc tttccttcaa gaagctcaga 780
taatgaaaaa attaagacat gataaacttg ttccactata tgctgttgtt tctgaagaac 840
caatttacat tgtcactgaa tttatgtcaa aaggtgctta ttccctttct attcgtgatt 900
gggatgagat aaggggtgac aatgtgaaac actacaaaat taggaaactt gacaatggtg 960
gatactatat cacaaccaga gcacaatttg atactctgca gaaattggtg aaacactaca 1020
cagaacatgc tgatggttta tgccacaagt tgacaactgt gtgtccaact gtgaaacctc 1080
agactcaagg tctagcaaaa gatgcttggg aaatccctcg agaatctttg cgactagagg 1140
ttaaactagg acaaggatgt ttcggcgaag tgtggatggg aacatggaat ggaaccacga 1200
aagtagcaat caaaacacta aaaccaggta caatgatgcc agaagctttc cttcaagaag 1260
ctcagataat gaaaaaatta agacatgata aacttgttcc actatatgct gttgtttctg 1320
aagaaccaat ttacattgtc actgaattta tgtcaaaaga gctccacaga cgctgcacag 1380
ctgtgaactc aactccagcc ctcagggaag gcagctcgca aagacctagg gcagcccgcc 1440
gatttcccaa aaagagcact gagaacccag caagcgtttt tctatatgat gctggcggca 1500
gcccagtgag ggaagtgcaa caagtatggt tctcctgctt gtcaactgtg gaaacagcga 1560
ccctgaaagt ggaggagcca cagctggaat cgtgttcgtc tgaatacacg cacctttcct 1620
atgagccttg taaagccagt cgttga 1646

<210> 95
<211> 415
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (109)..(170)
<223> a, c, g or t

<220>
<221> unsure
<222> (323)
<223> a, c, g or t

<220>
<221> unsure
<222> (325)
<223> a, c, g or t

<400> 95
gagagtgttt tagcccagaa tagtggaacc ctcaggcaaa gtttataaaa attaatacca 60
ctgcagaaag tgattaaata ccatggactg taggtttagg tttctgctnn nnnnnnnnnn 120
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnngtgcata 180
atggagattt taatagtagc catctcataa ggtggttgca aaggttaaat gtgttaatat 240
gcatgatgca catagaacaa tgcctagcac atagtagaga tacataatca ctactatata 300
ctggtaccag tananggtca ggtcttatgg acctaaggtc atataactta gtctcttcca 360
agattcttga aatgatttct caaaacaaga gaatataaag aagaaacgtt atgaa 415

<210> 96
<211> 504
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> (212)..(231)
<223> a, c, g or t

<400> 96
atatcctaat atgagcatgc ccattattcc tctgtcctat gaaacaaaaa tggcattttt 60
caatggattt gttttggata tataattagt tcatttgctg tttagaagcc ttgccaaaag 120
tgtttagatt ttggtactgc aactgctttc ctcttgccca gaaatgtttt gcctcttctt 180
ttcctacaag ttaaatgttc taaatataaa gnnnnnnnnn nnnnnnnnnn naattctaat 240
gtgaaaggca ctagctgtct aataggtttc atgtatcatt actattacta tatgtatctt 300
aatgtagtct atgtaggttt ttatcagaaa gtgtaccttt ctatggttta ttattttata 360
ttctggggcc ttttatctca gatataaacc atgaacagta atgatagtcc ctgacatata 420
aatcttagta aaaagtgatt aaaaatctaa aactcagtat gaaaaacata tcttgttagc 480
ataaattaaa accttttatt gttt 504

<210> 97
<211> 516
<212> DNA
<213> Homo sapiens

<400> 97
atatcctaat atgagcatgc ccattattcc tctgtcctat gaaacaaaaa tggcattttt 60
caatggattt gttttggata tataattagt tcatttgctg tttagaagcc ttgccaaaag 120
tgtttagatt ttggtactgc aactgctttc ctcttgccca gaaatgtttt gcctcttctt 180
ttcctacaag ttaaatgttc taaatataaa ggggtatgtg tgtgtgtgtg taattctaat 240
gtgaaaggca ctagctgtct aataggtttc atgtatcatt actattacta tatgtatctt 300
aatgtagtct atgtaggttt ttatcagaaa gtgtaccttt ctatggttta ttattttata 360
ttctggggcc ttttatctca gatataaacc atgaacagta atgatagtcc ctgacatata 420
aatcttagta aaaagtgatt aaaaatctaa aactcagtat gaaaaacata tcttgttagc 480
ataaattaaa accttttatt gtttaaaaaa aaaaaa 516

<210> 98
<211> 400
<212> DNA
<213> Homo sapiens

<400> 98
aattagatct ttcctgcaat aaggatctag gtggaggttt tgaagactcg ccggctcagt 60
tggtcatgct aaagcatcta caagtcctag atcttcacca gtgctcacta acagcagatg 120
acgtgatgtc actgacccag gtcattcctt tactttcaaa tcttcaagaa ttggatttat 180

cagccaacaa aaagatgggc agttcttctg aaaacttact cagcaggctc cgatttttac 240
cagcattgaa gtcattagtt atcaacaact gtgctttgga gagtgagact tttacagctc 300
ttgctgaagc ctctgttcac ctctctgctc tggaagtatt caacctttct tgggaacaag 360
tgtgttggtg ggcaacttga agctgcttct gggaaacact 400

<210> 99
<211> 2352
<212> DNA
<213> Homo sapiens

<400> 99
attttgattg aacacaggct tgacagaatc ttcttttctt cttagaaatc ctagaaaaca 60
gaaagcaaca ggaagatgtc ttattgggaa ctaccccat caacttcacc atgagtcaaa 120
caaggaagaa aacttcctca gaaggagaaa ctaagcccca gacttcaact gtcaacaaat 180
ttctcagggg ctccaatgct gaaagcagaa aagaggacaa tgaccttaaa acaagtgatt 240
cccaacccag cgactggata cagaagacag ccacctcaga gactgctaag cctctcagtt 300
cagaaatgga atggagatcc agtatggaga aaaatgagca tttcctgcag aagctgggca 360
aaaaggctgt caacaagtgt ctagatttga ataactgtgg attaacaaca gcggacatga 420
aagaaatggg agaagcattt gagatgattc ctgaacttga agagctaaat ttgtcttgga 480
acagtaaagt gggaggaaat ttgcctctga tccttcagaa gttccaaaaa gggagcaaga 540
tacaaatgat tgagcttgtg gcttgctccc tcacgtcaga agatgggaca tttctgggtc 600
aactgctacc tatgctgcaa agtctcgaag tacttgatct ttccattaac agagacattg 660
ttggcagtct gaacagtatt gctcagggat taaaaagcac ctcaaatctg aaagtactga 720
agttacattc atgtggatta tcacaaaaga gtgtcaaaat attggatgct gcttttaggt 780
atttgggtga gctgaggaaa ttagatcttt cctgcaataa ggatctaggt ggaggttttg 840
aagactcgcc ggctcagttg gtcatgctaa agcatctaca agtcctagat cttcaccagt 900
gctcactaac agcagatgac gtgatgtcac tgacccaggt cattccttta ctttcaaatc 960
ttcaagaatt ggatttatca gccaacaaaa agatgggcag ttcttctgaa aacttactca 1020
gcaggctccg atttttacca gcattgaagt cattagttat caacaactgt gctttggaga 1080
gtgagacttt tacagctctt gctgaagcct ctgttcacct ctctgctctg gaagtattca 1140
acctttcttg gaacaagtgt gttggtggca acttgaagct gcttctggaa acactaaagc 1200
tttccatgtc tcttcaagtg ctgaggctga gcagctgttc cctggtgaca gaggatgtgg 1260
ctctcctggc atcggtcata cagacgggtc atctggccaa actgcaaaag ctggacctga 1320
gctacaatga cagcatctgt gatgcggggt ggaccatgtt ctgccaaaac gtgcggttcc 1380
tcaaagagct aatcgagctg gatattagcc ttcgaccatc aaattttcga gattgtggac 1440
aatggtttag acacttgtta tatgctgtga ccaagcttcc tcagatcact gagataggaa 1500
tgaaaagatg gattctccca gcttcacagg aggaagaact agaatgcttt gaccaagata 1560
aaaaaagaag cattcacttt gaccatggtg ggtttcagta aactgatttc ccatgtccta 1620
ctaagctaca aaccattctc caaaggaaaa gaacatgaac gaattccaga gtcatgaact 1680
gaatttcaac ttctgggcca tttaatggga cttatattac aagagctttg taaatatata 1740
tatatattac atatatatat gtaatataca tatatacaca tatatataat atacatatat 1800
aatacacata tatatgtaaa tatatatata atatctaata tgagcatgcc attattctct 1860
gtctatgaaa caaaaatggc atttttcaat ggatttgttt tggatatata attagttcat 1920
ttgctgttta gaagccttgc caaaagtgtt tagattttgg tactgcaact gctttcctct 1980
tgcccagaaa tgttttgcct cttcttttcc tacaagttaa atgttctaaa tataaagggg 2040
tatgtgtgtg tgtgtgtaat tctaatgtga aaggcactag ctgtctaata gtttcatgta 2100
tcattactat tactatatgt atcttaatgt agtctatgta ggtttttatc agaaagtgta 2160
cctttctatg gtttattatt ttatattctg gtgcctttta tctcagatat aaaccatgaa 2220
cagtaatgat agtcactgac atataaatct tagtaaaaag tgattaaaaa tctaaaactc 2280
agtatgaaaa acatatcttg ttagaataaa ttaaaacctt ttattgttta aaaaattgtt 2340
aaaaaaaaaa aa 2352

<210> 100
<211> 565
<212> DNA
<213> Homo sapiens

<400> 100
atttgattg aacacaggct tgacagaatc ttcttttctt cttagaaatc ctagaaaaca 60
gaaagcaaca ggaagatgtc ttattgggaa ctacccccat caacttcacc atgagtcaaa 120
caaggaagaa aacttcctca gaaggagaaa ctaagcccca gacttcaact gtcaacaaat 180
ttctcagggg ctccaatgct gaaagcagaa aagaggacaa tgaccttaaa acaagtgatt 240
cccaacccag cgactggata cagaagacag ccacctcaga gactgctaag cctctcagtt 300
cagaaatgga atggagatcc agtatggaga aaaatgagca tttcctgcag aagctgggca 360
aaaaggctgt caacaagtgt ctagatttga ataactgtgg attaacaaca gcggacatga 420
aagaaatggt tgccttgctg ccttttctcc cagacttgga agaactggat atctcctgga 480
atggtttgta ggtggaaccc tcctttccat cactcagcaa atgcatctgg tcagcaagtt 540
aaaaatcttg aggctgggta gctgc 565

<210> 101
<211> 13
<212> PRT
<213> Homo sapiens

<400> 101
Met Leu Leu His Asp Ile Asp Trp His Leu Met Ser Ile
1 5 10

<210> 102
<211> 14
<212> PRT
<213> Homo sapiens

<400> 102
Met Val Leu Pro Gly Ser Leu Ser Met Leu Thr Tyr Gly Met
1 5 10

<210> 103
<211> 23
<212> PRT
<213> Homo sapiens

<400> 103
Met Gln Val Leu Tyr Trp Thr Tyr Leu Leu Leu Ile Leu Phe Pro Thr
1 5 10 15

Phe Thr Cys Leu Phe Ile Phe
20

<210> 104
<211> 26
<212> PRT
<213> Homo sapiens

<400> 104
Met Asn Leu Tyr Met Asn Leu Pro Ser Ala Val Arg Phe Ser Arg Ala
1 5 10 15

Thr Pro Leu Ile Ser Leu Phe Leu Ala Leu
20 25

<210> 105
<211> 49
<212> PRT
<213> Homo sapiens

<400> 105
Met Thr Thr Lys Lys Gln Glu Glu Cys Glu Ser Leu Lys Asp Lys Gln
1 5 10 15

Lys Ala Thr Lys Gln Ser Ile Ser Phe Cys Ile Tyr Ile Ile Lys Val
20 25 30

Lys Phe Ser Thr Leu Ala Thr Asp Tyr Lys Ser Val Pro Ser Gly Cys
35 40 45

Cys

<210> 106
<211> 61
<212> PRT
<213> Homo sapiens

<400> 106
Met Pro Ser Pro Ser Ala Pro Ser Ile Val Pro Val Leu His Gly Cys
1 5 10 15

Trp Val His Ile Cys Gln Ala Asp Val Tyr His Thr Leu Leu Lys Gly
20 25 30

Phe Lys Ser Val Phe Glu Thr Glu Ser His Val Val Ser Pro Arg Leu
35 40 45

Glu Cys Asn Gln Ser Lys Thr Pro Leu Lys Lys Asn Lys
50 55 60

<210> 107
<211> 34
<212> PRT
<213> Homo sapiens

<400> 107
Met Glu Leu Val Met Glu Trp Lys Leu Thr Ile Cys Ser Pro Lys Cys
1 5 10 15

Ala Thr Thr Thr Gln Gly Leu Gln Thr Asp Ser Tyr Leu Asp Val Val
20 25 30

Glu Ser

<210> 108
<211> 77
<212> PRT
<213> Homo sapiens

<400> 108

Met Val Asn Pro Ala Gln Glu Met Thr Leu Ser Arg Asn Thr Cys Lys
1   5   10   15

Tyr Lys Lys Gln Asp Ile Leu Pro Gln Leu Arg Ser Asp Lys Ile Thr
            20   25   30

Leu Gly Lys Leu Gln Gly Gln Cys Ala Ser Lys Thr Lys Ser Leu Val
        35   40   45

Ser Ser Leu Thr Ser Tyr Leu Pro Ala Phe Ile Ile Ile Ser Leu Ser
    50   55   60

Val Thr Gln Tyr Leu Val Asn Phe Leu Phe Trp His Thr
65   70   75

<210> 109
<211> 59
<212> PRT
<213> Homo sapiens

<400> 109

Met Gln Cys Lys His Phe Phe Leu Thr Tyr Leu Thr Asp Gln Gly Gly
1   5   10   15

Gln Val Ala Leu Leu Ser Ser Phe Pro Pro Cys Gly Asp Ser Gly Ile
            20   25   30

Gln Ala His Ser Ile Thr Arg Leu Ser His Ile Gly Val Phe His Phe
        35   40   45

Gly Asp Glu Asp Glu Gly Glu Ser Gly Arg Glu
    50   55

<210> 110
<211> 91
<212> PRT
<213> Homo sapiens

<400> 110

Met Asp Val Met Gly Lys Leu Lys Gly Ser Cys Asp Glu Thr Gly Ser
1   5   10   15

Glu Asn Ser Asp Gly Asp Leu Ser Lys Val Ile Leu Pro Lys His His
            20   25   30

Leu Ala Ile Met Ile Pro Pro Asn Leu Ser Gln Phe Val Tyr Phe Ile
        35   40   45

Ser Arg Gly Ser Phe Ser Val Leu Ala Ser Cys Val Phe Val Phe Phe
    50   55   60

Phe Phe Ser Val Ile Leu Gln Ala Gln Asp Phe Leu Leu Asp Thr Gly
65   70   75   80

Arg Ile Ser Leu Leu Lys Glu Ala Gly Gly Thr
                85   90

<210> 111
<211> 45
<212> PRT
<213> Homo sapiens

<400> 111
Met Gly His Val Asp Gln Leu Ser Pro Arg Thr Thr Asn Leu Ala Cys
1 5 10 15

Ser Asp Asp Leu Cys Ser Arg Gln Gly Phe Arg Leu Asp Cys Cys Ser
20 25 30

Ser Leu Trp Arg His Asn Pro Asn Cys Glu Leu Leu Asn
35 40 45

<210> 112
<211> 64
<212> PRT
<213> Homo sapiens

<400> 112
Met Leu Lys Met Ile Leu Ala Ser Ile Val Ile Asn Ser Val Ile Pro
1 5 10 15

Glu Phe Phe Val Ser Pro Arg His Thr Asn Phe Cys Pro Leu Leu Leu
20 25 30

Phe Ser Gln Ser Phe Leu Leu Ala Phe Leu Ser Asn Arg Val Leu Leu
35 40 45

Thr Pro Tyr Ile Pro Phe Trp Leu Val Arg Val Ser Phe Ser Ser Ser
50 55 60

<210> 113
<211> 25
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (14)

<220>
<221> UNSURE
<222> (17)

<400> 113
Met Leu Leu Phe Thr Lys Leu Leu Ile Ile Met Val Ile Xaa Ile Asn
1 5 10 15

Xaa Asn Asn Lys Leu Leu Gln Leu Phe
20 25

<210> 114
<211> 57
<212> PRT
<213> Homo sapiens

<400> 114
Met Arg Ile Gln Asn Leu Thr Cys Leu Leu Leu Gly Ser Lys Glu Met
1 5 10 15

Ser Thr Ser Ser Pro Leu Thr Pro Asn Gly Val Glu Gly Phe Gly Pro
20 25 30

Gln His Cys Val Thr Tyr Ser His His Asp Phe Leu Ala Gln Val Thr
35 40 45

Pro Ser Val Lys Trp Lys Arg Glu Glu
50 55

<210> 115
<211> 147
<212> PRT
<213> Homo sapiens

<400> 115
Met Asn Glu Ser Trp Ala Gly Pro Gly Pro Ala Glu Arg Ala Glu Glu
1 5 10 15

Ala Val Ser Gly Val Gly Val Glu Ala Lys Thr Gln His Ala Gly Gln
20 25 30

Gly Ala Gln Pro Gly Gly Met Gly Cys Gly Phe Ser Ser Gly Pro Ile
35 40 45

Gly Met Ala Leu Gly Leu Gly Leu Val Gly Thr Ala Ala Thr Arg Gly
50 55 60

Gly Ser Ser Ala Trp Pro Asp Ser Thr Cys Asn Val Gly Arg Gln Trp
65 70 75 80

Ala Pro Pro Gly Gly Arg Asn Thr Val Arg Ser Met Gln Arg Ala Gly
85 90 95

Asp His Gly Ala Cys Asp Leu Arg Ala His Pro Gly Gln Thr Trp Val
100 105 110

Arg Gly Gly Leu Gly Arg Gln Asp Ser Glu Gly Leu Gln Gly Val Phe
115 120 125

Val Leu Cys Pro Tyr Thr Gly Asp Leu His Gly Arg Val Arg Ser Ile
130 135 140

Arg Met Leu
145

<210> 116

<211> 73
<212> PRT
<213> Homo sapiens

<400> 116
Met Thr Ile Ser Leu Cys Ala Thr Asn Leu Pro Arg Ala Ala Thr Val
1   5   10   15

Leu Arg Met Lys Pro Lys Leu Pro Gly Ser Gly Pro Val Gln His Glu
            20   25   30

Pro His Leu Pro Ser Gln Pro Gln His Pro Leu Leu Phe Phe Gln Ala
        35   40   45

Gly Gly Lys Leu Glu Ala His Pro His Phe Thr Gln Thr Leu Gly Ile
    50   55   60

Pro Ile Ser Gly Asn Arg Gly Val Phe
65   70

<210> 117
<211> 48
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (46)

<400> 117
Met Tyr Asn Ile Leu Lys Ala Phe Asp Lys Ile Val His Ile Ile Ser
1   5   10   15

Asn Thr Ile Leu Tyr Tyr Tyr Gln Gln His Lys Ala Asn Val Ser Lys
            20   25   30

Asn Ser Arg Leu Arg Ile Ser Lys Asn Ser Pro Arg Ala Xaa Phe Arg
        35   40   45

<210> 118
<211> 38
<212> PRT
<213> Homo sapiens

<400> 118
Met Leu Pro Val Ser Pro Thr Leu Lys Glu Arg Asn Gln Arg Arg Met
1   5   10   15

Leu Leu Lys Ser Thr His Leu Ala Ser Val Ser Ser Ala Ser Cys Thr
            20   25   30

Gln Thr Lys His Thr Gly
        35

<210> 119
<211> 55
<212> PRT
<213> Homo sapiens

<400> 119
Met Lys Ile Phe Ile Ile Ile Leu Ser Pro Leu Cys Gly Ile Leu Leu
1 5 10 15

Asn Val Leu Glu Ser Leu Lys Phe Ile Phe Lys Cys Glu Ser Leu Leu
20 25 30

Phe Val Trp Gly Glu Glu Cys Gln Val Gly Ile Met Asn Gln Ala Leu
35 40 45

Pro Tyr Gln Val Leu Leu Tyr
50 55

<210> 120
<211> 92
<212> PRT
<213> Homo sapiens

<400> 120
Glu Ser His Thr Leu Gln Val Ile Leu Gly Cys Glu Met Gln Glu Asp
1 5 10 15

Asn Ser Thr Glu Gly Tyr Trp Lys Tyr Gly Tyr Asp Gly Gln Asp His
20 25 30

Leu Glu Phe Cys Pro Asp Thr Leu Asp Trp Arg Ala Ala Glu Pro Arg
35 40 45

Ala Trp Pro Thr Lys Leu Glu Trp Glu Arg His Lys Ile Arg Ala Arg
50 55 60

Gln Asn Arg Ala Tyr Leu Glu Arg Asp Cys Pro Ala Gln Leu Gln Gln
65 70 75 80

Leu Leu Glu Leu Gly Arg Gly Val Leu Asp Gln Gln
85 90

<210> 121
<211> 85
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (51)..(72)

<400> 121
Met Ile Lys Val Ser Leu Thr Ser Ala Pro Lys Val Ser Ser Leu Glu
1 5 10 15

Gly Thr Asn Arg Arg Glu His Ser Asp Thr Gln Gly Pro Leu Ser Val

```
            20                  25                  30

Pro Trp Lys Pro Ser Asp Leu Cys Arg Pro Ile Ser Val Arg Lys Trp
        35                  40                  45

Val Ala Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
    50                  55                  60

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Arg Thr Thr Gln Ser Ser Trp Gln
65                  70                  75                  80

Ile Leu Asn Lys Gly
                85


<210> 122
<211> 20
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (15)

<400> 122
Met Gly Gly Ala Trp Ser Ile Ala Gly Pro Leu Thr Gly Phe Xaa Phe
  1               5                  10                  15

Arg Leu Thr Phe
            20


<210> 123
<211> 103
<212> PRT
<213> Homo sapiens

<400> 123
Phe Tyr Phe Leu Phe Ser Phe Val Leu Arg Trp Ser Phe Thr Leu Val
  1               5                  10                  15

Thr Gln Ala Gly Val Gln Trp Cys Asp Leu Gly Ser Leu Gln Pro Pro
            20                  25                  30

Pro Pro Arg Leu Lys Ala Phe Ser Cys Leu Gly Leu Pro Ser Ser Trp
        35                  40                  45

Asp Tyr Arg His Ala Leu Gln Arg Pro Ala Asn Phe Ala Phe Leu Val
    50                  55                  60

Glu Ile Gly Phe His His Val Gly Gln Ala Gly Pro Gln Leu Leu Thr
65                  70                  75                  80

Ser Gly Asp Pro Ser Ile Leu Ala Ser Gln Ser Ala Gly Ile Thr Gly
                85                  90                  95

Val Thr Ala Val Pro Gly Pro
            100
```

<210> 124
<211> 48
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (13)..(43)

<400> 124
Met Val Val Ile Gln Ala Xaa Glu Glu Glu Lys Thr Xaa Xaa Xaa Xaa
1 5 10 15

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
20 25 30

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Ile Trp Lys Ile Cys
35 40 45

<210> 125
<211> 95
<212> PRT
<213> Homo sapiens

<400> 125
Met Ser Ser Tyr Met Ile Asn Lys Phe Leu Pro Ile Lys Lys Val Lys
1 5 10 15

Ile Pro Gly His Lys Val Phe Ser Thr Asp Ile Met Phe Leu Lys Phe
20 25 30

Val Ser Ile Ala Thr Leu Leu Arg Arg His Thr Asp Ile Ser Glu Asp
35 40 45

Leu Arg Val Leu Gln Asn Thr Glu Lys Ile Ser Arg Arg Lys Gly Lys
50 55 60

Gly Glu Thr Lys Lys Leu Lys Glu Gly Leu Thr Tyr Lys Trp Asn Asp
65 70 75 80

Leu Lys Arg Asn Gly Glu Pro Gly Glu Thr Gly Val Ser Gln Ser
85 90 95

<210> 126
<211> 48
<212> PRT
<213> Homo sapiens

<400> 126
Met Ile Lys Tyr Phe Lys Ser Asn Asn Tyr Lys Phe Asn Tyr Tyr Lys
1 5 10 15

Thr Ser Ser Leu Thr Ser Asp Cys Phe Val Leu Ser Phe Lys Ile Ile

20 25 30

Met Val Cys Leu Arg Val Cys Leu Leu Asn Thr Phe Ala Tyr Leu Pro
35 40 45

<210> 127
<211> 98
<212> PRT
<213> Homo sapiens

<400> 127
Met Glu Phe Arg Ser Val Ala Gln Val Gly Val Gln Trp Arg Asp Leu
1 5 10 15

Gly Leu Leu Gln Pro Leu Pro Leu Gln Phe Lys Gln Phe Tyr Cys Leu
20 25 30

Ser Leu Ser Ser Ser Trp Asp Tyr Arg His Ser Pro Pro His Pro Ala
35 40 45

Asn Phe Leu Tyr Phe Ala Lys Ile Leu Tyr Ile Ala Lys Arg Phe His
50 55 60

His Val Gly Gln Ala Gly Leu Ala Leu Leu Thr Ser Gly Asp Pro Pro
65 70 75 80

Thr Ser Ala Ser Gln Ser Ala Gly Ile Thr Gly Leu Ser His Cys Ala
85 90 95

Gln Pro

<210> 128
<211> 50
<212> PRT
<213> Homo sapiens

<400> 128
Met Gly Lys Arg Arg Asp Ser Trp Thr Asn Arg Glu Arg Gln Leu Glu
1 5 10 15

Asn Lys Ser Met Gln Lys Ile Ile Tyr Asn Lys Ile Met His Leu Thr
20 25 30

Leu Val Thr Lys Gln Ile Ser Tyr Pro His Phe Ser Leu Ser Val Phe
35 40 45

Val Ser
50

<210> 129
<211> 16
<212> PRT

<213> Homo sapiens

<400> 129
Met Leu Leu Phe Val Leu Ser Leu Val Phe Gln Tyr Gln Phe Asn Thr
1   5   10   15

<210> 130
<211> 54
<212> PRT
<213> Homo sapiens

<400> 130
Met Ala Leu His Cys Phe Thr Ser Gly Leu Trp Ile Ala Ser Val Arg
1   5   10   15

Lys Lys Val Lys Met Lys Glu Lys Val Glu Gln Ile Leu Ala Thr Glu
            20   25   30

Pro Pro Glu Asp Ser Cys Pro Phe Ser Asn Lys Leu Ser Gly Lys Cys
        35   40   45

Cys Cys His Gly Ser Thr
    50

<210> 131
<211> 41
<212> PRT
<213> Homo sapiens

<400> 131
Met Cys Ala His Lys Gly Lys Ala Met Arg Glu Arg Thr Gln Pro Glu
1   5   10   15

Gly Gly His Leu Ala Ser Gln Gly Glu Ala Leu Arg Glu Thr Lys Pro
            20   25   30

Ala Arg Leu Gly Thr Val Ala His Gly
        35   40

<210> 132
<211> 35
<212> PRT
<213> Homo sapiens

<400> 132
Met Ala Leu Ile Leu Leu Glu Ala Leu Cys Phe Gly Leu Ile Ile Cys
1   5   10   15

Met Asn Arg Glu Ser Ile Ser Thr Leu Ile Phe Tyr Lys His Trp Met
            20   25   30

Ser Ile Leu
        35

<210> 133

<211> 58
<212> PRT
<213> Homo sapiens

<400> 133
Met Phe Asn Ala Tyr Leu Leu Tyr Asn Asn Gln Val Ile Thr Val Gln
1 5 10 15

Ile Lys Gly Pro Lys Cys Phe Arg Tyr Asp Ile Ile Leu Ser Ile Val
20 25 30

Asn Trp Thr Lys Glu Thr Leu Tyr Val Gln Gly Ser Val Glu Gln Pro
35 40 45

Trp Cys Ser Trp Asp Met Leu Pro Arg Cys
50 55

<210> 134
<211> 27
<212> PRT
<213> Homo sapiens

<400> 134
Met Met Lys Leu Cys Phe Thr Ala Ser Leu Leu His Gly Ala Leu Leu
1 5 10 15

Trp His Leu Ala Thr Thr Asn Ser Leu Ile Pro
20 25

<210> 135
<211> 46
<212> PRT
<213> Homo sapiens

<400> 135
Met Glu Leu Pro Ser Met Cys Pro Ile Leu Phe Phe Val Thr Val Phe
1 5 10 15

Phe Met Tyr His Thr Pro Ser Cys Pro Ser Ser Val Pro Gln Thr His
20 25 30

Gln Ser His Phe Leu Leu Thr Ala Leu Gly Leu Ala Leu Thr
35 40 45

<210> 136
<211> 77
<212> PRT
<213> Homo sapiens

<400> 136
Met Thr Cys Pro Gly Gly Glu Thr Gly Trp Gly Cys Leu Arg Met Asp
1 5 10 15

Pro Arg Glu Trp Val Ser Ser Pro Asp Gln Gln Asn Leu Arg Met Cys
20 25 30

Ala Trp Ile Gln Pro His Leu Lys Leu Gly Leu His Phe Val Ser Gly
35 40 45

Ala Pro Asn Ala Leu Cys Leu Gly Cys Leu Tyr Ser Trp His Thr Gly
50 55 60

Glu Ala Leu Ser Pro Ala Gly Pro Gly Cys Cys Cys Ser
65 70 75

<210> 137
<211> 37
<212> PRT
<213> Homo sapiens

<400> 137
Met Glu Gln Glu Ser Val Pro Ser Met Ser Leu Phe Thr Arg Ile Leu
1 5 10 15

Ser Gln Pro Ser Leu Phe Pro Trp Gln Ala Leu His Arg Glu Thr Gly
20 25 30

Lys Arg Ser Thr Val
35

<210> 138
<211> 59
<212> PRT
<213> Homo sapiens

<400> 138
Met Leu Leu Pro Leu Pro Ala Ile Ser Phe Pro Cys Asn Ser Leu Phe
1 5 10 15

His Pro Ala Asp Ala Ser Ser Leu Ser Trp Leu Ser Ser Lys Ser Tyr
20 25 30

Pro Leu Gly Lys Leu Thr Arg Met Leu Gln Ser Asp Gly Val Ser Pro
35 40 45

Pro Gly Pro Pro Gln Thr Leu Tyr Phe Leu Leu
50 55

<210> 139
<211> 50
<212> PRT
<213> Homo sapiens

<400> 139
Met Asp Asn Lys Cys Leu Thr Leu Thr Asn Tyr Leu Ala Ile Met Gly
1 5 10 15

Phe Phe Asp Gln Lys Ser Ser Lys Arg Val Trp Trp Gly Leu Arg Asp
20 25 30

Pro Ser Ser Leu Pro Lys Asn Met Lys Ser Phe His Phe Gln Tyr Val
35 40 45

```
Lys Thr
    50


<210> 140
<211> 72
<212> PRT
<213> Homo sapiens

<400> 140
Met Arg Val Val Phe Lys Ile Thr Phe Cys Arg Val Val Cys Ser Thr
 1               5                  10                  15

Leu Met Leu Lys Gly Ser His Leu Pro Gln Pro Ile Lys Leu Cys Cys
            20                  25                  30

Leu Cys Ser Ala Phe Tyr His Lys Asn Met Thr Phe Lys His Lys Asn
        35                  40                  45

Thr Leu Tyr Ser Thr Thr Lys Asn Arg Asn Asp Ile Tyr Leu His Cys
    50                  55                  60

Phe Pro Ile Ser Leu His Leu Tyr
65                  70


<210> 141
<211> 863
<212> PRT
<213> Homo sapiens

<400> 141
Met Pro Glu Gln His Lys Asp Pro Arg Val Gln Glu Asn Pro Asp Asp
 1               5                  10                  15

Gln Arg Thr Val Pro Glu Val Thr Gly Asp Ala Arg Ser Ala Phe Trp
            20                  25                  30

Pro Leu Arg Asp Asn Gly Gly Pro Ser Pro Phe Val Pro Arg Pro Gly
        35                  40                  45

Pro Leu Gln Thr Asp Leu His Ala Gln Ser Ser Glu Ile Arg Tyr Asn
    50                  55                  60

His Thr Ser Gln Thr Ser Trp Thr Ser Ser Ser Thr Lys Arg Asn Ala
65                  70                  75                  80

Ile Ser Ser Ser Tyr Ser Ser Thr Gly Gly Leu Pro Gly Leu Lys Gln
                85                  90                  95

Arg Arg Gly Pro Ala Ser Ser Arg Cys Gln Leu Thr Leu Ser Tyr Ser
            100                 105                 110

Lys Thr Val Ser Glu Asp Arg Pro Gln Ala Val Ser Ser Gly His Thr
        115                 120                 125

Arg Cys Glu Lys Gly Ala Asp Thr Ser Pro Gly Gln Thr Ile Ala Pro
    130                 135                 140
```

```
Thr Gly Gly Ser Pro Arg Ser His Asp Ser Arg Pro Arg Arg Arg Lys
145                 150                 155                 160

Ile Pro Leu Leu Pro Arg Arg Arg Gly Glu Pro Leu Met Leu Pro Pro
                165                 170                 175

Pro Leu Glu Leu Gly Tyr Arg Val Thr Ala Glu Asp Leu His Leu Glu
            180                 185                 190

Lys Glu Thr Ala Phe Gln Arg Ile Asn Ser Ala Leu His Val Glu Asp
        195                 200                 205

Lys Ala Ile Pro Asp Cys Arg Pro Ser Arg Pro Ser His Thr Leu Ser
    210                 215                 220

Ser Leu Ala Thr Gly Ala Ser Gly Gly Pro Pro Val Ser Lys Ala Pro
225                 230                 235                 240

Thr Met Asp Ala Gln Gln Asp Arg Pro Lys Ser Gln Asp Cys Leu Gly
                245                 250                 255

Leu Val Ala Pro Leu Ala Ser Ala Ala Glu Val Pro Ala Thr Ala Pro
            260                 265                 270

Val Ser Gly Lys Lys His Arg Pro Pro Gly Pro Leu Phe Ser Ser Ser
        275                 280                 285

Asp Pro Leu Pro Ala Asn Ser Ser His Ser Arg Asp Ser Ala Gln Val
    290                 295                 300

Thr Ser Met Ile Pro Ala Pro Phe Thr Ala Ala Ser Arg Asp Ala Gly
305                 310                 315                 320

Met Arg Arg Thr Arg Ser Ala Pro Ala Ala Ala Ala Ala Ala Pro Pro
                325                 330                 335

Pro Ser Thr Leu Asn Pro Thr Ser Gly Ser Leu Leu Asn Ala Val Asp
            340                 345                 350

Gly Gly Pro Ser His Phe Leu Ala Ser Ala Thr Ala Ala Ala Arg Ala
        355                 360                 365

Gln Arg Ser Glu Val Arg Tyr Asn Gln Arg Ser Gln Thr Ser Arg Thr
    370                 375                 380

Arg Ser Cys Leu Lys Arg Asn Ala Ser Ser Ser Ser His Ser Ser Thr
385                 390                 395                 400

Glu Gly Leu Gln Glu Val Lys Arg Arg Arg Gly Pro Ala Ser Ser His
                405                 410                 415

Cys Gln Leu Ala His Ser Ser Ser Asn Thr Val Ser Glu Asp Gly Pro
            420                 425                 430

Gln Ala Val Ser Ser Gly His Arg Cys Glu Asn Lys Ala Gly Thr Ala
        435                 440                 445

Pro Gly Gln Thr Leu Ala Pro Arg Gly Gly Ser Pro Arg Ser Gln Ala
```

450 455 460

Ser Arg Pro His Ile Asn Thr Ala Leu His Val Glu Asp Lys Ala Ile
465 470 475 480

Ser Asp Cys Arg Pro Ser Arg Pro Ser His Thr Leu Ser Ser Leu Ala
485 490 495

Thr Gly Ala Ser Gly Gly Pro Pro Val Ser Lys Ala Pro Thr Met Asp
500 505 510

Ala Gln Gln Asp Arg Pro Lys Ser Gln Asp Ser Leu Gly Leu Leu Ala
515 520 525

Pro Leu Ala Ser Ala Ala Glu Val Pro Ser Thr Ala Pro Val Ser Gly
530 535 540

Lys Lys His Arg Pro Pro Gly Pro Leu Phe Ser Ser Ser Asp Pro Leu
545 550 555 560

Pro Ala Thr Ser Tyr His Ser Arg Asp Thr Ala Gln Val Thr Ser Leu
565 570 575

Ile Pro Ala Thr Phe Thr Ala Ala Ser Arg Asp Ala Gly Met Arg Arg
580 585 590

Thr Arg Ser Ala Pro Ala Ala Ala Thr Ala Ala Pro Pro Pro Ser Thr
595 600 605

Leu Asn Asn Thr Ser Gly Ser Leu Leu Asn Ala Val Asp Gly Gly Pro
610 615 620

Ser His Phe Leu Ala Ser Ala Thr Ala Ala Ala Arg Ala Gln Arg Ser
625 630 635 640

Glu Val Arg Tyr Asn Gln Arg Ser Gln Thr Ser Arg Thr Arg Ser Cys
645 650 655

Leu Lys Arg Asn Ala Ser Ser Ser Ser Ser Ser His Ser Ser Thr Glu
660 665 670

Gly Leu Gln Glu Val Lys Arg Arg Arg Gly Pro Ala Ser Ser His Cys
675 680 685

Gln Leu Ala His Ser Ser Ser Asn Thr Val Ser Glu Asp Gly Pro Gln
690 695 700

Ala Val Ser Ser Gly His Arg Cys Glu Asn Lys Ala Gly Thr Ala Pro
705 710 715 720

Gly Gln Thr Leu Ala Pro Arg Gly Gly Ser Pro Arg Ser Gln Ala Ser
725 730 735

Arg Pro His Ile Asn Ser Ala Leu His Val Glu Asp Lys Ala Ile Ser
740 745 750

Asp Cys Arg Pro Ser Arg Pro Ser His Thr Leu Ser Ser Leu Ala Thr
755 760 765

Gly Ala Ser Gly Gly Pro Pro Val Ser Lys Ala Pro Thr Met Asp Ala
770 775 780

Gln Gln Asp Arg Pro Lys Ser Gln Asp Cys Leu Gly Leu Leu Ala Pro
785 790 795 800

Leu Ala Ser Ala Ala Glu Val Phe Ser Thr Ala Pro Val Ser Gly Lys
805 810 815

Lys His Arg Pro Pro Gly Pro Leu Phe Ser Ser Ser Asp Pro Leu Pro
820 825 830

Ala Thr Ser Ser His Ser Gly Asp Ser Ala Gln Asp Thr Ser Leu Ile
835 840 845

Pro Ala Pro Phe Thr Pro Ala Ser Arg Asp Ala Gly Ile Arg Arg
850 855 860

<210> 142
<211> 29
<212> PRT
<213> Homo sapiens

<400> 142
Met Ser Tyr Leu Ser Leu Leu Leu Ile Ser Ile Phe Met Val Cys Tyr
1 5 10 15

Phe Lys Arg Asn Ser Phe Pro Ile Thr Ile Leu Phe Ser
20 25

<210> 143
<211> 32
<212> PRT
<213> Homo sapiens

<400> 143
Met Pro Trp Pro Met Pro Ile Cys Thr Gly Thr Gln Gly Val Leu Thr
1 5 10 15

His Arg Gln Gly Pro Pro Pro Ala Ala Val Gly Val Ser Pro His Thr
20 25 30

<210> 144
<211> 29
<212> PRT
<213> Homo sapiens

<400> 144
Met Asn Ala Phe Leu Leu Glu Arg Met Thr Glu Ser Gln Ala Met Asp
1 5 10 15

Ile Gln Thr Cys Ile Phe Gln Thr Leu Leu Glu Asn Lys
20 25

<210> 145
<211> 48
<212> PRT
<213> Homo sapiens

<400> 145
Met Ile Val Thr Asn Thr Ile Leu Lys Phe Ile His Lys Lys Pro Thr
1   5   10   15

Thr Ile Thr Pro Thr Lys Gln His Gly Ile Ile Phe Ser Val Pro Glu
            20   25   30

Ala Lys Val Arg Ala Leu Leu Cys Phe Leu Leu Arg Met Pro Ser Pro
        35   40   45

<210> 146
<211> 55
<212> PRT
<213> Homo sapiens

<400> 146
Gly Gln Ala Leu Trp Leu Met Pro Val Ile Pro Val Val Ala Lys Ala
1   5   10   15

Glu Gly Lys Asp His Leu Arg Pro Gly Val Ala Asn Gln Pro Gly Gln
            20   25   30

His Ser Lys Thr Leu Phe Leu Gln Lys Lys Asn Phe Ala Lys Leu Ala
        35   40   45

Glu His Gly Gly Ala Cys Leu
    50   55

<210> 147
<211> 55
<212> PRT
<213> Homo sapiens

<400> 147
Met Ser Arg Phe Arg Ile Gln Thr Ser Glu Thr Ala Pro Ile Pro Leu
1   5   10   15

Val Ser His Pro His Thr Pro Leu Ser Asn Asn Asn Asn Leu His Leu
            20   25   30

Gly Asn Val Cys Tyr Val Pro Gly His Thr Gly Ile Ile Ser Cys Thr
        35   40   45

Pro His Arg His Leu Ile Lys
    50   55

<210> 148
<211> 50
<212> PRT
<213> Homo sapiens

<400> 148
Met Gln Gly Leu His Leu Pro Gln Gly Leu Gly Thr Cys Tyr Ser Ile
1   5   10   15

Cys Leu Gln Cys Leu Ser Pro His Gly Tyr Phe Phe Val Ala Val Gly
            20   25   30

Leu Ser Ser Asn Val Met Ser Pro Thr Ser Leu Pro Lys Ala Val Leu
        35   40   45

Pro Thr
    50

<210> 149
<211> 31
<212> PRT
<213> Homo sapiens

<400> 149
Met Leu Pro Val Asn Ile Ser His Pro Leu Ser Arg Gly Asn Pro Leu
1   5   10   15

Leu Ser Ser Lys Phe Ser Lys Phe Phe Leu Ile Glu Phe Ser Gln
            20   25   30

<210> 150
<211> 36
<212> PRT
<213> Homo sapiens

<400> 150
Met Asp Tyr Ser Leu Ser Phe Asp Asn Tyr Thr Trp Gly Phe Gly Glu
1   5   10   15

Pro Arg Ile Lys Val Gln Ser Phe Asn Asp Leu Leu Ala Pro Gly Leu
            20   25   30

Thr Gln Glu His
        35

<210> 151
<211> 85
<212> PRT
<213> Homo sapiens

<400> 151
Met Ile Arg Ser Lys Gly Thr Asn Phe Gln Ile Leu Ala Glu Leu Phe
1   5   10   15

Lys Gly Met Asp Phe Leu Trp Leu Gln Leu Ala Arg Leu Phe Gln Lys
            20   25   30

Ala Cys Pro Trp Leu Thr Ala Cys Leu Ala Gln Phe Leu Arg Ser Pro
35 40 45

Leu Val Met Glu Asn Arg Ala Asp Arg Ile Gln Met Ala Arg Phe His
50 55 60

Arg Gly Gln Gly Gly Pro Gln Ser Ala Asn Gln Gly Arg Leu Arg Pro
65 70 75 80

Glu Lys Gly Ile Ser
85

<210> 152
<211> 73
<212> PRT
<213> Homo sapiens

<400> 152
Met Asp Arg Phe Leu Asn Ser Lys Ala Arg Arg Leu Gly Ser Cys Ser
1 5 10 15

His Pro Ala Phe Tyr Leu Leu Cys Val Pro Asp Glu Asp Thr Ser Cys
20 25 30

Ser Thr Met Tyr Leu Pro Leu Lys Arg Arg Ala Asp Pro Asp Gln Leu
35 40 45

Phe Ser Asp Leu Leu Gly Gly Thr Gln Arg Leu Trp Arg Leu Trp Pro
50 55 60

Ser Leu Ala Ser Val Glu Ser Gly Leu
65 70

<210> 153
<211> 63
<212> PRT
<213> Homo sapiens

<400> 153
Met Gln Cys Thr Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10 15

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
20 25 30

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Lys Ile Lys Phe Gly
35 40 45

Met Lys Gln Glu Leu Ser Trp Thr Ile Tyr Asn Leu Leu Arg Tyr
50 55 60

<210> 154
<211> 46
<212> PRT
<213> Homo sapiens

<400> 154
Met Arg Cys Leu Leu Ala Asp Ser Ser Leu Gln Met Gln Pro Gly Asp
1 5 10 15

Val Thr Leu Arg Leu Glu Ser Cys Gly Ser Asn Pro Arg Gln Arg Gln
20 25 30

Leu His Gln Val Leu Val Trp Val Arg Asn Arg Gly Lys Gly
35 40 45

<210> 155
<211> 72
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (22)

<400> 155
Met Pro Pro Arg Gly Trp Ala Cys Pro Ser Ser Gly Pro Pro Ala Pro
1 5 10 15

Gly Pro Gly Arg Trp Xaa Arg Ala Ala Ala Gly Gly Leu Arg Arg Thr
20 25 30

Arg Cys Asp Trp Leu Pro Leu Arg Arg Thr Gln Met Ser Leu Arg Arg
35 40 45

Ile Asp Leu Leu Pro Ser Pro Ala Gly Gln Ala Gln Ala Gly Ser Glu
50 55 60

Asn Tyr Leu Pro Leu Phe Ile Ser
65 70

<210> 156
<211> 20
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (10)

<220>
<221> UNSURE
<222> (13)..(14)

<220>
<221> UNSURE
<222> (16)

<220>
<221> UNSURE
<222> (20)

<400> 156
Met Val Phe Ile Phe Ser Thr Thr Ile Xaa Phe Phe Xaa Xaa Glu Xaa
1 5 10 15

Glu Ser Cys Xaa
20

<210> 157
<211> 66
<212> PRT
<213> Homo sapiens

<400> 157
Met Ser Leu Thr Tyr Ser Trp Lys Lys Ser Lys Val Thr Lys Phe Asn
1 5 10 15

Leu Ser Thr Leu Arg Met Thr Val Thr Asn Lys Asn Arg Thr Val Gln
20 25 30

Lys Cys Ala Lys Asp Thr Arg Lys Leu Asn Asn Ile Asn Ser Met Ile
35 40 45

Ile Val Ile Leu Tyr Thr Met Glu Ser Lys Gln Ile Phe Phe His Gly
50 55 60

Asn Ser
65

<210> 158
<211> 41
<212> PRT
<213> Homo sapiens

<400> 158
Met Met Thr Gly Glu Ala Arg Glu Ser Gln Ile Ala Leu Tyr Lys Gln
1 5 10 15

Arg Phe Arg Glu Phe Arg Glu Glu Gly Arg Thr Ile Tyr Lys Gly Arg
20 25 30

Trp Lys Arg Ser His Leu Ala Glu Gly
35 40

<210> 159
<211> 31
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (7)

<400> 159
Met Leu Glu Leu Gly Leu Xaa Pro Lys Leu Thr Ser Glu Tyr Arg Phe
1 5 10 15

Pro Pro Asn Cys Met Ile Leu His Ile Trp Ser Gln Leu Glu Val
20 25 30

<210> 160
<211> 75
<212> PRT
<213> Homo sapiens

<400> 160
Met Tyr Ile Tyr Ile Cys His His Phe Lys Asn Gln Ala Phe Lys Val
1 5 10 15

Lys Leu Ser Phe Pro His Ile Phe Phe His Ser Ile Phe Tyr Gln Tyr
20 25 30

Arg His Ser Leu Leu Leu Leu Ser Phe Gln Phe Pro Ile Ile Ser Ser
35 40 45

His Pro Ile Phe Cys Ala Ser Ser Val Phe Lys Thr His Ser Pro Ser
50 55 60

Ala Ala Met Ala Pro Thr Ile Ile Phe Ile Thr
65 70 75

<210> 161
<211> 36
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (7)..(13)

<400> 161
Met Lys Arg Gly Asn Leu Xaa Xaa Xaa Xaa Xaa Xaa Xaa Gly Thr Pro
1 5 10 15

Cys Lys Asp Trp Ser His Thr Ala Met Ser Gln Glu Pro Pro Val Leu
20 25 30

Val Arg Val Leu
35

<210> 162
<211> 24
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (9)

<220>
<221> UNSURE
<222> (20)

<400> 162
Met Trp Ala Ala Trp Arg Arg Arg Xaa Asn Gly Phe Phe Pro Arg Ile
1 5 10 15

Pro Gly Lys Xaa Arg Gly Pro Asn
20

<210> 163
<211> 31
<212> PRT
<213> Homo sapiens

<400> 163
Met Cys His Ser Leu Tyr Arg Phe Leu Asn Cys His Ser Arg Tyr Tyr
1 5 10 15

Ile Val Tyr Thr Tyr Leu Thr Ile Phe Tyr Trp Cys His His Phe
20 25 30

<210> 164
<211> 134
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (2)..(22)

<220>
<221> UNSURE
<222> (39)..(67)

<220>
<221> UNSURE
<222> (79)..(113)

<400> 164
Met Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10 15

Xaa Xaa Xaa Xaa Xaa Xaa Ala Gly Lys Arg Glu Asn Gln Lys Asp Ser
20 25 30

Ser Val Arg Arg Thr Trp Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
35 40 45

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
50 55 60

Xaa Xaa Xaa Arg Phe Ser Pro Arg Ala Tyr Arg Lys Lys Val Xaa Xaa
65 70 75 80

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
85 90 95

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
100 105 110

Xaa Arg His Asn Arg Lys Leu Ile His Leu Ser Ser Lys Phe Leu Ile
115 120 125

Ile Asn Val Ile Ala Ser
130

<210> 165
<211> 51
<212> PRT
<213> Homo sapiens

<400> 165
Met Ser Lys Val Asp Leu Phe Ile Thr Asp Ser Phe Lys Lys Phe Asn
1 5 10 15

Gln Tyr Leu Leu Ala Thr Tyr Ser Thr Ser Gly Thr Gln Gly Ile Trp
20 25 30

Ser Thr Thr Met Lys Lys Arg Asp Trp Thr Leu Lys Glu His Arg Ser
35 40 45

Cys His Phe
50

<210> 166
<211> 60
<212> PRT
<213> Homo sapiens

<400> 166
Met Ser Asp Ser Arg Leu Cys Ser Cys Phe Leu His Thr Leu Ile Phe
1 5 10 15

Leu Asn Ile Ser Lys Ile Gln Ser Gly Ser Lys Ile Thr Cys Lys Asn
20 25 30

Ile Leu Ala Gln Glu Phe Asp Arg Leu Lys Ile Asn Tyr Leu Lys Tyr
35 40 45

Ile Lys Gln Glu Val Tyr Leu Leu Tyr Ser Met Tyr
50 55 60

<210> 167
<211> 15
<212> PRT
<213> Homo sapiens

<400> 167
Met Val Phe Gln Lys Thr Ser Leu Tyr Ser Asn Asn Ile Leu Leu
1 5 10 15

<210> 168
<211> 106
<212> PRT

<213> Homo sapiens

<400> 168
Cys Pro Ala Ala Tyr Thr Val Phe Leu Thr Arg Ile Ile Val Lys Tyr
1 5 10 15

Tyr Leu Asn Arg Gly Leu Phe Ser Glu Thr Pro Ser Asn Leu Lys Val
20 25 30

Glu Glu Lys Gly Trp Val Trp Trp Leu Met Pro Val Thr Pro Ala Leu
35 40 45

Trp Glu Ala Glu Ala Gly Gly Ser Leu Glu Leu Ser Leu Arg Pro Gly
50 55 60

Trp Ala Thr Pro Ser Leu Pro Lys Asn Thr Lys Met Ser Gln Ala Trp
65 70 75 80

Trp Cys Thr Pro Val Val Pro Ala Ala Leu Gly Ala Glu Val Gly Gly
85 90 95

Arg Leu Gly Pro Arg Arg Trp Arg Leu Gln
100 105

<210> 169
<211> 19
<212> PRT
<213> Homo sapiens

<400> 169
Met Gly Pro Asp Arg Leu Lys Gln Lys Ser Asn Thr Ala Val Val Ser
1 5 10 15

Arg Trp Ile

<210> 170
<211> 47
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (3)..(4)

<220>
<221> UNSURE
<222> (13)

<220>
<221> UNSURE
<222> (16)

<400> 170
Met Asp Xaa Xaa Lys Trp Arg Met Arg Arg Gln Pro Xaa Ile Asn Xaa
1 5 10 15

Met Tyr Gln Thr Val Thr Ile Cys Glu Glu Tyr Cys Val Tyr Thr Asn
20 25 30

Arg Lys Gln Leu Lys Ala Phe Asn Met Cys Gly Trp Gly Glu Arg
35 40 45

<210> 171
<211> 197
<212> PRT
<213> Homo sapiens

<400> 171
Gln Glu Ala Gln Ile Met Lys Lys Leu Arg His Asp Lys Leu Val Pro
1 5 10 15

Leu Tyr Ala Val Val Ser Glu Glu Pro Ile Tyr Ile Val Thr Glu Phe
20 25 30

Met Ser Lys Gly Ala Tyr Ser Leu Ser Ile Arg Asp Trp Asp Glu Ile
35 40 45

Arg Gly Asp Asn Val Lys His Tyr Lys Ile Arg Lys Leu Asp Asn Gly
50 55 60

Gly Tyr Tyr Ile Thr Thr Arg Ala Gln Phe Asp Thr Leu Gln Lys Leu
65 70 75 80

Val Lys His Tyr Thr Glu His Ala Asp Gly Leu Cys His Lys Leu Thr
85 90 95

Thr Val Cys Pro Thr Val Lys Pro Gln Thr Gln Gly Leu Ala Lys Asp
100 105 110

Ala Trp Glu Ile Pro Arg Glu Ser Leu Arg Leu Glu Val Lys Leu Gly
115 120 125

Gln Gly Cys Phe Gly Glu Val Trp Met Gly Thr Trp Asn Gly Thr Thr
130 135 140

Lys Val Ala Ile Lys Thr Leu Lys Pro Gly Thr Met Met Pro Glu Ala
145 150 155 160

Phe Leu Gln Glu Ala Gln Ile Met Lys Lys Leu Arg His Asp Lys Leu
165 170 175

Val Pro Leu Tyr Ala Val Val Ser Glu Glu Pro Ile Tyr Ile Val Thr
180 185 190

Glu Phe Met Ser Lys
195

<210> 172
<211> 59
<212> PRT
<213> Homo sapiens

<220>

<221> UNSURE
<222> (28)..(49)

<400> 172
Met Cys Ile Met His Ile Asn Thr Phe Asn Leu Cys Asn His Leu Met
1 5 10 15

Arg Trp Leu Leu Leu Lys Ser Pro Leu Cys Thr Xaa Xaa Xaa Xaa Xaa
20 25 30

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
35 40 45

Xaa Gln Lys Pro Lys Pro Thr Val His Gly Ile
50 55

<210> 173
<211> 56
<212> PRT
<213> Homo sapiens

<220>
<221> UNSURE
<222> (14)..(21)

<400> 173
Met Lys Pro Ile Arg Gln Leu Val Pro Phe Thr Leu Glu Xaa Xaa Xaa
1 5 10 15

Xaa Xaa Xaa Xaa Xaa Leu Tyr Leu Glu His Leu Thr Cys Arg Lys Arg
20 25 30

Arg Gly Lys Thr Phe Leu Gly Lys Arg Lys Ala Val Ala Val Pro Lys
35 40 45

Ser Lys His Phe Trp Gln Gly Phe
50 55

<210> 174
<211> 104
<212> PRT
<213> Homo sapiens

<400> 174
Met Leu Lys His Leu Gln Val Leu Asp Leu His Gln Cys Ser Leu Thr
1 5 10 15

Ala Asp Asp Val Met Ser Leu Thr Gln Val Ile Pro Leu Leu Ser Asn
20 25 30

Leu Gln Glu Leu Asp Leu Ser Ala Asn Lys Lys Met Gly Ser Ser Ser
35 40 45

Glu Asn Leu Leu Ser Arg Leu Arg Phe Leu Pro Ala Leu Lys Ser Leu
50 55 60

Val Ile Asn Asn Cys Ala Leu Glu Ser Glu Thr Phe Thr Ala Leu Ala